モジュール型学習機構を用いたマルチエージェント環境における競合行動の同時学習

Simultaneous Learning to Acquire Competitive Behaviors in Multi-Agent System based on Modular Learning System

○ 野間 健太郎 (阪大)　正 高橋 泰岳 (阪大, 阪大 FRC)　正 浅田 稔 (阪大, 阪大 FRC)

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

The existing reinforcement learning approaches have been suffering from the policy alternation of others in multiagent dynamic environments. A typical example is a case of RoboCup competitions since other agent behaviors may cause sudden changes in state transition probabilities of which constancy is needed for the learning to converge. The keys for simultaneous learning to acquire competitive behaviors in such an environment are

- a modular learning system for adaptation to the policy alternation of others, and
- an introduction of macro actions for simultaneous learning to reduce the search space.

This paper presents a method of modular learning in a multiagent environment, by which the learning agents can simultaneously learn their behaviors and adapt themselves to the situations as consequences of the others' behaviors.

Key Words: reinforcement learning, competitive behaviors acquisition, multi-agent system, modular learning system, simultaneous learning, macro actions, and RoboCup

1 緒言

エージェントが複数存在するマルチエージェント環境に強化学習を適用し，複数のエージェントがそれぞれの行動を同時に学習する研究が多くなされている [1)2)3)5)]．適切な行動が獲得されるためには，学習者から見た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仮定が必要である．しかし，ロボカップのような環境では他のエージェントの行動政策の変化により学習者の視点から見た状態遷移確率が大きく変化し，単一の学習器では適切な行動を獲得することは困難であった．

Ikenoue et al[3)] は学習中は学習者の政策を固定にすることにより協調行動の獲得を実現している．この手法では，それぞれの学習者の行動政策を比較的長い間固定とすることで，学習者の視点から見た状態遷移確率がほぼ一定とみなすことができ，目的とする行動を獲得することができた．マルチエージェント環境下での協調行動の獲得の場合，お互いの利益が一致するため，一度協調行動を獲得してしまえばその行動を取り続けることで目的を達成できる傾向がある．しかし競合行動の獲得の場合，お互いの利益が相反するため自分自身の成功は相手にとっては失敗となり，相手の成功は自分自身にとっては失敗となる．失敗したエージェントは行動価値を低く見積もり，他の行動をとるようになる．一方，成功したエージェントは相手の政策変化により，同じ行動をとり続けると失敗する．このように常に各エージェントの政策は更新され，その結果，学習者から見た状態遷移が一定とならず，学習が非常に困難なものとなる．

Jacobs and Jordan et al[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 [7)] は，環境の予測性に基づいて非線形/非定常な環境を時空間的に分割し，予測を正しく行なっているモジュールが制御を行なう MOSAIC を提案した．単純な予測器が並列に状態遷移を予測し，その予測の最も良い予測器と対となる制御器が責任を持って環境を制御し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．枝澤 [6)] はマルチエージェント環境における競合行動の獲得において相手の政策を一定にし，その状況に状態遷移モデルと行動を学習しやすい環境を作りこむために，学習スケジューリングを行ない，学習者の立場から見た状態遷移が一定と見なせるような状況を各学習器に割り当てた．相手の政策の変化に応じて学習器を切替え，その学習器に従った行動をとることで状況に応じた行動を獲得できた．しかし，複数のエージェントが非同期で独立に学習する場合は，スケジューリングを行なうことができず，各学習器に状態遷移が一定となるような状況を割り当ててやることが難しい．Stefan et al[2)] はマクロ行動を導入することにより，2 台のロボットの協調行動の獲得を実時間で実現している．マクロ行動の導入により，探索空間を抑え，状態遷移を素早く獲得し，また相手に自分の行動を観測させやすくさせた．

そこで本稿では，複数の学習器を用いマクロ行動を導入することで，マルチエージェント環境下における競合行動の同時学習を獲得する手法を提案し，Robocup 中型機リーグに出場しているロボットに適用することでその有効性を示す．

2 A Multi-Module Learning System

各学習器は，予測器 (predictor) と計画器 (planner) からなる (図 1)．予測器は環境の相互作用のみからなる状態遷移モデルを持ち，計画器はその状態遷移モデルに基づいて，動的計画法の枠組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遷移確率を基に，現在の状

況を最も良く予測している学習器を選択し，その計画器にしたがった行動をとる．

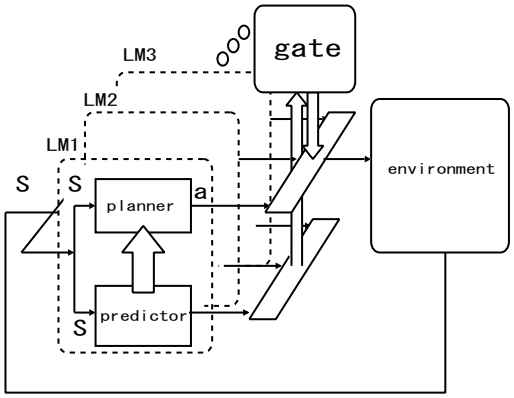


Fig.1 A multi-module learning system

2.1 予測器

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

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

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はなくある状態 s と行動 a が与えられたときの次の報酬の期待値

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

を推定するモデルを持つ．

2.2 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 $\hat{\mathcal{R}}^a_{ss'}$ が決まるとある状態 s，行動 a における行動価値関数 $Q(s,a)$ は，

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

のように与えられる．この値を基に動的計画法の枠組で計画する．ここで γ は減衰係数を表す．

2.3 学習器の選択

ある状況 (状態 s で行動 a をとり次状態 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式 (4) のように予測される．各モジュールでの信頼度 g_i(式 (5)) は，ある一定期間 T の間の p_t を考慮し，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をとることができる．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．

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

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

2.4 学習器の更新

ある状態遷移を状態遷移モデルに入れる時，過去と未来の一定期間の T の間の p_t を考慮し，信頼度の大きな学習器にデータを入れる．その結果，現在の状況に対して最適な学習器を更新することができる．

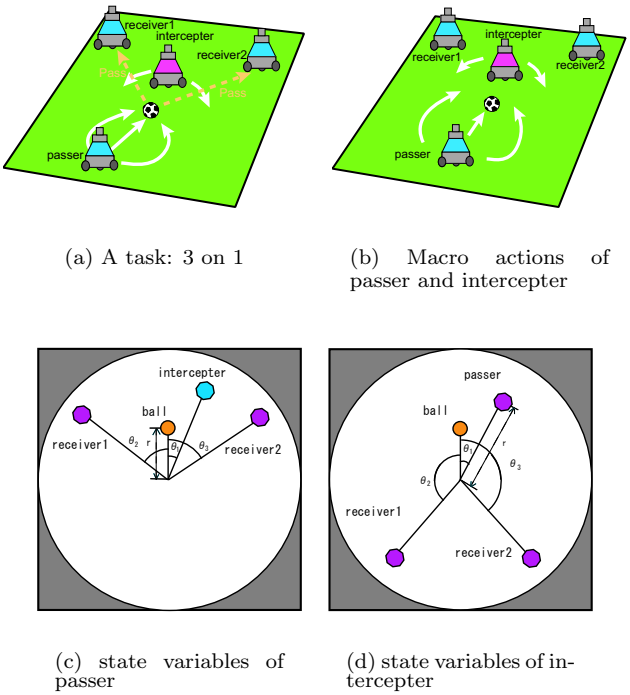
$$g_t = \prod_{t=t-T}^{t+T} \exp(\lambda p_t) \tag{6}$$

3 Task and assumption

本手法を RoboCup 中型機リーグに出場しているサッカーロボットを用いたタスクに適用する．2 台の receiver のどちらかにパスを行なう passer と receiver へのパスをインターセプトする intercepter がいる 3 on 1 のような状況を考える (図 2(a))．intercepter が片方の receiver のパスコースに入った場合，passer はもう片方の receiver にパスをするようにボールまでのアプローチを学習する．intercepter がどちらの receiver のパスコースに入るかという政策の変化から，その状況にふさわしい学習器に切り替えて，その学習器の計画器で計画された行動をとることにより，intercepter の政策の変化に対応する．また，intercepter は passer の動きに応じて，receiver の間のパスコースに入り，パスをインターセプトすることを学習する．passer がどちらの receiver に向かってパスをするかという政策の変化から，その状況に適した学習器に切り替えて，その学習器の計画器で計画された行動をとることにより，passer の政策の変化に対応する．passer と intercepter がそれぞれ学習器 2 個を用い，状況に応じて学習器を切替えて同時に学習する．

3.1 状態/行動空間

passer の状態空間 S は，前方カメラ上でのボールの y 座標，全方位カメラ画像上における各ロボットとボールの角度である (図 2(c))．マクロ行動 A はボールを常に正面で見ながらボールの回りを右に回る，左に回る，まっすぐ近付くマクロ行動を与える (図 2(b))．intercepter の状態空間 S は，全方位カメラ画像上における passer の距離，各ロボットとボールの角度である (図 2(d))．マクロ行動 A はボールを常に正面で見ながら左または右側の receiver へのパスコースをブロックするという行動を与える (図 2(b))．



(a) A task: 3 on 1

(b) Macro actions of passer and intercepter

(c) state variables of passer

(d) state variables of intercepter

Fig.2 Task, macro actions, and state spaces of passer and intercepter

3.2 ゴール状態の判定

passer の各行動におけるゴール状態の判定は，ボールまで辿り着き，パスをした時，intercepter にパスカットされず receiver にパスが通れば報酬+1 を与える．intercepter

の各行動におけるゴール状態の判定は，passer がボールまで辿り着き，パスをした時，passer と receiver とのパスコースに入ってインターセプトすれば報酬+1 を与える．また，passer と intercepter が双方とも失敗する場合や試行がある一定時間を越えた場合は引き分けとし，双方に報酬 0 を与える．

4 実験結果

4.1 獲得された行動の様子

シミュレータにおいて同時学習により獲得されたパスの様子を図 3 に示す．また，simulation で得られたデータを実機に実装して獲得されたインターセプトの様子を図 4 に示す．

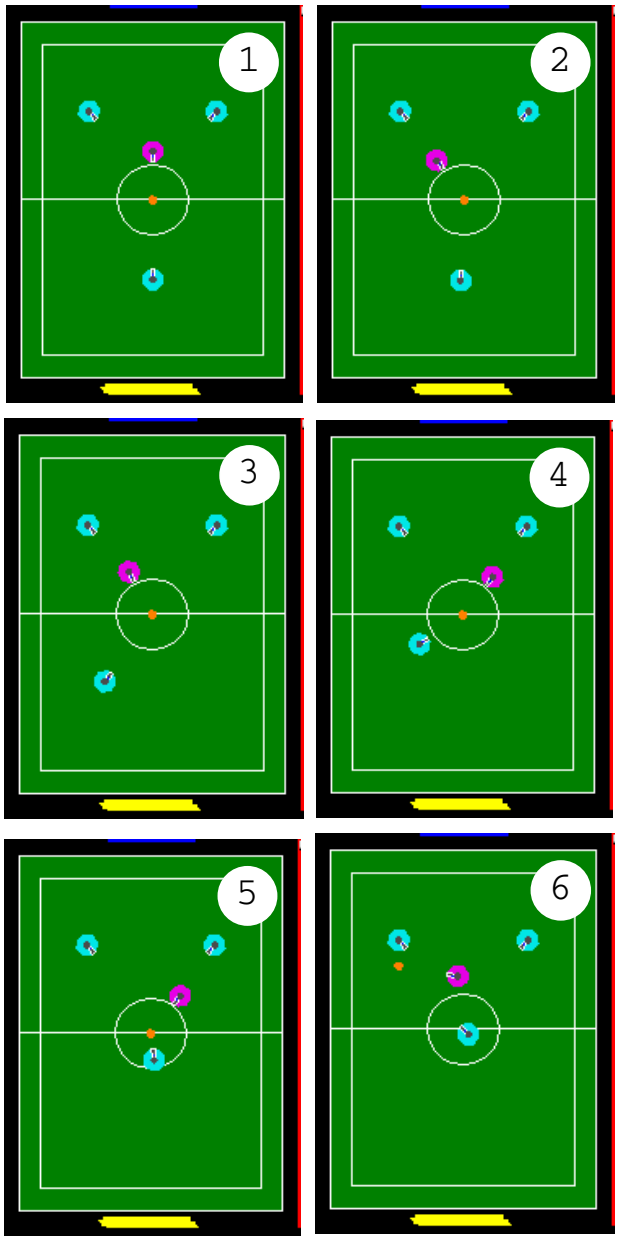


Fig.3 a sequence of a behavior of passing a ball to the left receiver in simulation

4.2 学習中のタスク成功率

シミュレーションによるタスク成功率を図 5 に示す．600 試行あたりまで intercepter の学習が進み，しだいに passer が intercepter に対応して学習し，1000 試行以降では両者ともほぼ同じ程度の成功率となっている．

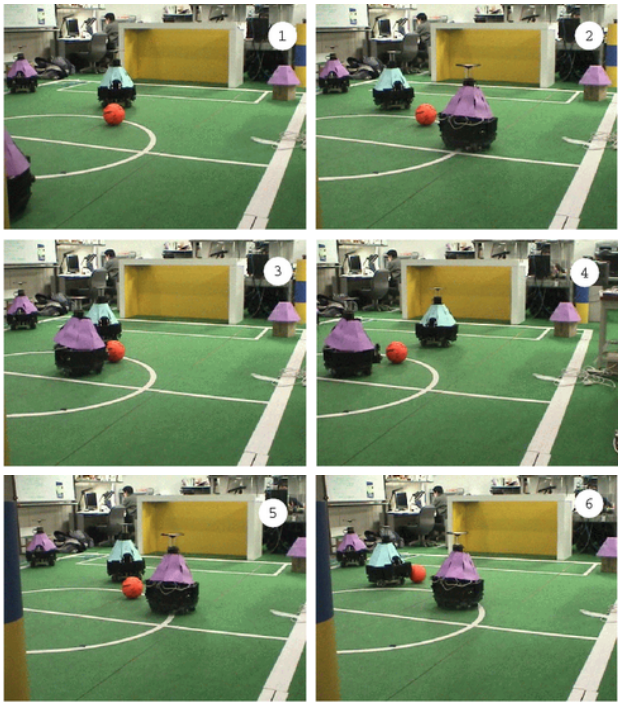


Fig.4 a sequence of a behavior of intercepting a pass to the left receiver in hardware

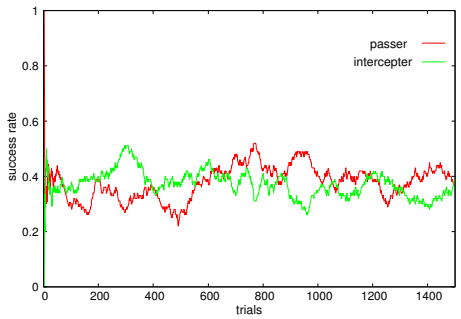


Fig.5 Curves of success rate in simulation

4.3 学習後の成功率の比較

シミュレーションで同時に学習させたあと，二つの学習器と単一の学習器を使った場合と二つの学習器を使ったエージェントと固定政策 (fixed) をしているエージェントの 300 試行の成功率を 表 1 に示す．LM は学習器を表している．0 または 1 は使用した学習器の ID を表している．draw rate とは，passer と intercepter が両方とも失敗した時や，一定時間が過ぎタイムオーバーとなった時である．学習器を 2 つ用いた方が状況に応じた行動がとれていることがわかる．

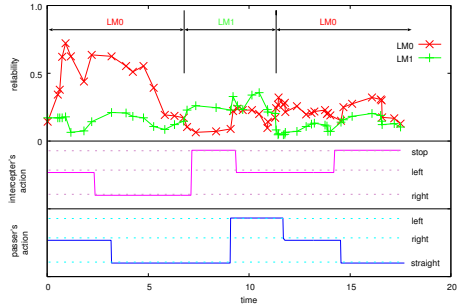
4.4 相手の政策による学習器の切替え

シミュレーションで同時学習後，同じ状況下で intercepter の政策を変化させた時の passer の信頼度の変化を図 6(a), 図 6(b) に示す．passer の right はボールの回りを右に回る (左側の receiver にパスをしようとする)，left は左に回る (右側の receiver にパスをしようとする)，straight は前に進む行動を表している．intercepter の right はボールの回りを右に回る (左側の receiver のパスコースを防ごうとする)，left は左に回る (右側の receiver のパスコースを防ごうとする)，stop は receiver とボールの間に入りパスコースを防いで状態を表している．intercepter の行動は同時学習中の行動を参考にして設計者によって決められている．11 秒付近まで passer も intercepter 同じ

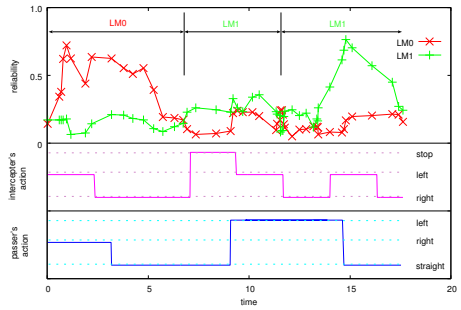
Table 1 success rate in simulation

passer	intercepter	passer's success rate[%]	intercepter's success rate [%]	draw rate[%]
LM0,LM1	LM0	59.0	23.0	18.0
LM0,LM1	LM1	52.7	34.3	13.0
LM0	LM0,LM1	25.6	55.0	19.4
LM1	LM0,LM1	26.0	59.3	14.7
LM0,LM1	fixed	84.0	9.3	6.7
fixed	LM0,LM1	2.0	91.7	6.3
LM0,LM1	LM0,LM1	37.6	37.3	25.1

行動をとり同じ状況にいる．situation1 では，intercepter は，その後も左に回る行動とり続け，situation2 では，右に回る行動に変化させた状況である．intercepter の政策の変化によって situation1 では，passer の LM1 の信頼度が下がり，LM0 の信頼度が上がっている．situation2 では，LM1 の信頼度が上がり LM0 の信頼度が下がっている．つまり，intercepter が右側の receiver のパスコースを防ごうとしている状況に対しては学習器 0 が担当し，左側の receiver のパスコースを防ごうとしている状況に対しては学習器 1 が担当していることになる．それぞれの状況で，信頼度が大きな学習器の計画器に従った行動をとることで 12 秒付近で passer の行動が situation1 では右に回りその後左の receiver にパスを成功させ，situation2 では左に回りその後，右側の receiver にパスを成功させている．相手の政策の変化による状況変化に対し，その状況に適した学習器を選択し，その計画器に従った行動をとることにより，その状況に適した行動がとれていることがわかる．



(a) situation1



(b) situation2

Fig.6 The transition of passer's reliablility by change of intercepter's policy in simulation

5 おわりに

本稿ではマルチエージェント環境下で競合行動の同時学習を行ない，それぞれのエージェントが目的の行動を獲得したことを示した．従来の強化学習において他のエージェントの政策の変化により状態遷移確率が大きく変化するような環境では最適な行動が獲得できず再学習が必要であるという問題を，複数の学習器用い，マクロ行動を導入することで解決した．マクロ行動を導入することで，探索空間を抑え，状態遷移モデルを素早く獲得し，状態遷移が一定とみなせるような状況を各学習器に割り当てる．また，マクロ行動をとることで相手に自分の行動を観測させやすくさせた．その結果，相手の政策の変化に対し学習器の切替がうまくいき，その状況下での最適な行動をとることができることをサッカーロボットを用いたシミュレーションと実機による実験で示した．

参考文献

[1] 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.

[2] Stefan Elfwing, Eiji Uchibe, Kenji Doya, and Henrik I. Chirstensen. Multi-agent reinforcement learning: Using macro actions to learn a mating task. *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 13, pp. 3164–2220, 2004.

[3] Shoichi Ikenoue, Minoru Asada, and Koh Hosoda. Cooperative behavior acquisition by asynchronous policy renewal that enables simultaneous learning in multiagent environment. In *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] Peter Stone, Richard S.Sutton, and Gregory Kuhlmann. Scaling reinforcement learning toward robocup soccer. *Journal of Machine Learing Research*, Vol. 13, pp. 2201–2220, 2003.

[6] 枝澤一寛. 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. Master's thesis, 大阪大学大学院 工学研究科 知能・機能創成工学専攻, 2003.

[7] 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.