学術・技術論文

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獲得

高橋泰岳[*1] 枝澤一寛[*1] 野間健太郎[*1] 浅田稔[*1*2]

Acquisition of Competitive Behaviors in Multi-Agent System based on a Modular Learning System

Yasutake Takahashi[*1], Kazuhiro Edazawa[*1], Kentarou Noma[*1] *and* Minoru Asada[*1*2]

Existing reinforcement learning approaches have been suffering from policy alternation by others in multi-agent dynamic environments that may cause sudden changes in state transition probabilities of which constancy is needed for behavior learning to converge. A typical example is the case of RoboCup competitions because behaviors of other agents may change the state transition probabilities. A modular learning system would be able to solve this problem if we can assign each module to one situation in which the module can regard the state transition probabilities as constant. Scheduling for learning is introduced to avoid the complexity in autonomous situation assignment. Furthermore, introduction of macro actions reduces the exploration space and it would enable agents to learn competitive behaviors simulaneously in such an adversary environment. This paper presents a method of modular learning in a multi-agent environment in which the learning agents can learn their behaviors and adapt themselves to the resultant situations by the others' behaviors.

Key Words: Reinforcement Learning, Competitive Behaviors Acquisition, Multi-agent System, Modular Learning System, Simultaneous Learning, RoboCup

1. はじめに

近年，ロボット自身が環境との相互作用を通して行動を獲得する強化学習 [1] が注目され，実際のロボットに適用する研究が多くなされてきた [1–10]．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か，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．これは状態遷移確率が変化すると学習が収束しないからである．しかしながらマルチエージェント環境下では他のエージェントの制御方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可能性が大きく，目的の行動を獲得することは従来手法では一般に困難である．そこで多くは，エージェントから見て状態遷移確率がほぼ一定とみなせる環境下で行動学習している．例えば Mataric [11] は複数のロボットがパックを目的領域に搬送するタスクを同時に学習する例を示したが，このタスクでは自身の行動の結果に対して他者の行動の影響を受けることが少なく，また逆に個々の最適行動の探索が他者の学習に悪影響を及ぼさないと仮定している．Asada ら [12] はシステム同定手法を用い，学習者と他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，かつ他エージェントは固定方策をとる必要がある．そのため他エージェントの制御方策が変化する場合は適用できない．Ikenoue ら [13] は，それぞれの学習者の行動制御方策の切替を非常にゆっくり行うことにより，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制御方策が変化したときには再学習をする必要がある．Kuhlman と Stone [14] はロボカップシミュレーションリーグの環境下で 3 対 2 のパス回しの行動獲得を行わせた．ただし学習する主体はボールを保持しているエージェントのみで，他のレシーバや相手チームのエージェントは固定の制御方策で行動するため，学習するエージェントのから見れば環境の状態遷移は一定とみなせる．

マルチエージェント環境下では一般に他のエージェントの方策は変化するので，学習者はこれに対処する必要がある．他者の方策に応じて制御方策をそれぞれ学習できる枠組があれば，この問題を解決できる．Jacobs と Jordan [15] は複数の学習モジュールを用い，各学習モジュール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

原稿受付 2007 年 11 月 19 日
[*1]大阪大学大学院工学研究科
[*1]JST ERATO 浅田共創知能システムプロジェクト
[*1]Graduate School of Engineering, Osaka University
[*1]JST ERATO Asada Synergistic Intelligence Project

れる学習システムを提案している．各学習モジュールの状況に対する適応度を重みとすれば，広く適用できる [16–19]．同様の考え方で環境のダイナミクスの変動に対応可能な学習アルゴリズムも提案されている [20,21]．しかしこれらの研究では，環境から得られる観測データの予測については扱っているが，ロボットやエージェントの行動獲得のための探索行動とその影響については考慮されていない．鮫島ら [22] や Haruno ら [23] 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探索空間を時空間的に分割し，予測を正しく行うモジュールに制御を行わせるものである．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相手の方策によって状態遷移が動的に変化する場合は，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重要となる．

本論文では，他者の制御方策が動的に変動するマルチエージェント環境下で複数の学習モジュールを用い，相手の方策に対応した適切な行動獲得が可能な手法を提案し，モジュール型学習機構を用いたロボットの競合行動を獲得する際の挙動や有用性を示す．まず 2 章で想定するタスクと仮定を示す．次に 3 章で提案するモジュール型学習機構を説明する．4 章では，まず相手の方策変動を認識するための探索空間が大きすぎるため，モジュール型の学習機構の適用だけでは実現が困難であることを示す．この課題を解決するために学習スケジューリングが必要であることを示す．次に 5 章では，探索空間を絞りこむことで，学習スケジューリングなしに複数のエージェントが競合行動を同時に学習をすることが可能であることを示す．探索空間を絞りこむためマクロ行動 [9] を導入する．これにより，自律的に状態遷移モデルの割り当てを行い，またその学習モジュールの計画器を用いて行動することが可能になる．最後に 6 章で結言を述べる．

2. タスクと仮定

ここではロボカップの中型機リーグに出場しているサッカーロボットのタスクを想定する．ロボットは前方視覚，全方位視覚，全方向移動機構およびキック機構を備え，味方プレイヤへのパス行動を獲得させる（**Fig.1** 参照）．環境にはボールとパスを邪魔するインタセプタが 1 体，パサーとパスを受けるためのレシーバが 2 体存在する．この環境下でパサーはインタセプタのディフェンスの動きに応じて，適切なレシーバにパスするために必要なボールまでのアプローチ行動を学習し，パスを実現する．これに対しインタセプタはパスを阻止するための制御方策を学習する． この問題設定から他者の行動方策のバリエーションが少ないため，学習者は数個の学習器で相手の方策変動に対応可能な問題を扱う．なお，行動学習はパサーはボールを特定のレシーバの方向に蹴るためのアプローチ行動，インタセプタはパサーが蹴るパスを阻止するための行動を学習するため，パサーがボールを蹴るまでを 1 試行とし，ボールを蹴った後の状態遷移は学習しない．

Fig.2 および **Fig.3** に実際に使用する実ロボットとそのシミュレータを示す．各ロボットは色情報を使った単純な画像処理により周りのプレイヤおよびボールを毎秒 30 フレームで認識する．Fig.3 では Fig.1 の状況のシミュレーションを示す．

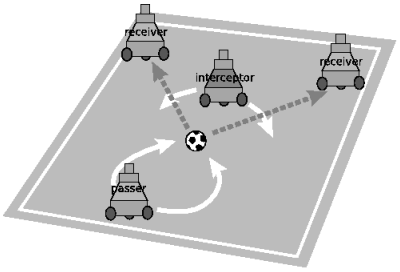


Fig. 1 A task: 3 on 1

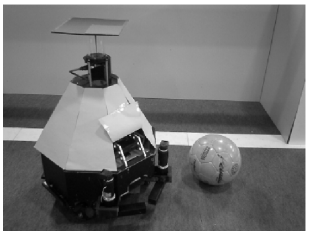


Fig. 2 A real robot

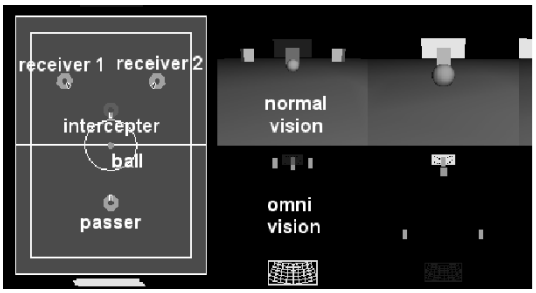


Fig. 3 Viewer of simulator

3. 複数学習モジュールによるマルチエージェント環境下における行動獲得

3.1 アプローチと仮定

強化学習の枠組では，システム全体を環境と学習者の相互作用としてモデル化する．学習者は自身のセンサを通して環境の現在の状態を認識し，環境に対して行動を出力する．それに応じて次時刻に環境は変化し，その結果報酬が学習者に与えられる．学習者が合目的的な行動を獲得するためには，学習者と環境との相互作用による状態遷移は学習中はマルコフ過程に従うと仮定する．マルチエージェント環境下での学習においては，環境のなかに他者が存在し，この他者が学習者に対する方策を変化させることで，学習者の目から見た環境下の状態遷移が一定には見えないことが問題となり，行動学習の収束を困難にする．

ここでは一定の状態遷移確率で遷移していく状況を単に状況と呼ぶ．そこで一つの状況に一つの学習モジュールを割り当て，学習モジュールが割り当てられた状況において合目的的な行動を獲得することを考える．また学習者は

- 環境のモデルを事前には持たず
- 状態遷移の観察によって，モデルを選択および同定し，
- 行動学習はモデルベース型強化学習の枠組で行う．

学習者はある状況において選択される各学習モジュールを，その予測の確からしさ（最も良い予測）から判断し，その学習モジュールを用いて行動学習する．

3.2 システム概要

提案システムを **Fig.4** に示す．実線で囲まれた各学習モジュールは予測器（predictor）と計画器（planner）を持つ．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．ゲートは各学習モジュールが予測する状態価値関数の値を基に，現在の状況を最もよく予測している学習モジュール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．

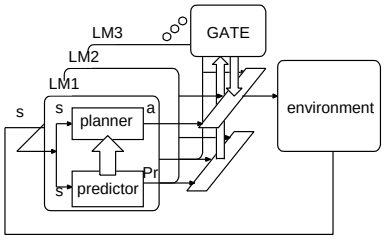


Fig. 4 A multi-module learning system

3.3 予測器

各学習モジュールの予測器は状態遷移モデルを持ち，ある状態 s で行動 a をとった時に次状態 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．このモデルは単純に学習者が環境の中で試行錯誤したときのデータをデータベースに残し，事後確率を計算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ち，そのモデルの獲得方法も同様である．状態遷移モデルと報酬モデルは状態遷移確率および報酬の期待値とも 0 で初期化され，経験があったところから更新される．

3.4 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 $\hat{\mathcal{R}}^a_s$ 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

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

で与えられる [24]．ここで γ は減衰係数を表す．動的計画法の枠組でこの値を計算し，最適な行動方策を得る．

3.5 行動選択のための学習モジュールの選択

行動選択のための学習モジュールの切り換えには，以下で定義する信頼度 ${}^e g_i$ を用いる．各モジュールの信頼度 ${}^e g_i$ 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モジュール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，各学習モジュールの予測モデルに基づいて

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

で予測される．この予測値を用いて，信頼度は

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

と計算する．ここで λ は適当なスケーリングパラメータであり，以降の実験では 0.2 で固定した．

3.6 モデル修正のための学習モジュールの選択と生成

モデル修正のための学習モジュールの選択についても同様の考え方に基づいて行う．ただし，ある時刻のデータがどの学習モジュールの状況にあっているかの判断は，その時刻以前のデータだけではなく，以後のデータも含めて判断する．

$${}^u g = \prod_{t=t-T}^{t+T} e^{\lambda p_t} \qquad (6)$$

この信頼度 ${}^u g$ の最大値を返す学習モジュールが経験したデータをつかってモデルを修正する．ただし，それまで経験したことのない状態と行動のペアからの次状態への状態遷移確率は計算できない．この状態遷移確率を低く見積もると，常に新しい学習モジュールにその経験が割り振られる．これを防ぐためにため，既存学習モジュールはその状態遷移確率を大きいもの（今回の実験では 60%）と推定することで，無駄な新しい学習モジュールの生成を抑制する．学習前は学習器を一つのみ用意する．この信頼度がある一定期間（1 秒），ある閾値（0.1）よりも低い場合に新しい学習器を追加し，この学習器に新たに状態遷移を学習させる．

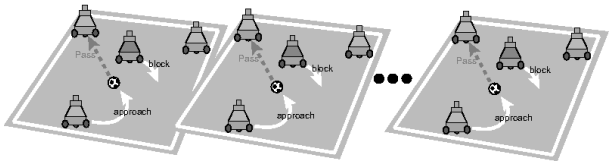
4. 行動学習におけるスケジューリングの必要性

前章の行動学習システムを 2 章で述べたタスクに適用する．状況の分別とそれに応じた行動獲得は膨大な探索空間のために困難を極める．一つの解決策として学習のスケジューリングが考えられる．例えば Asada ら [2] はゴール状態に到達しやすい状況から学習を始め，徐々にゴール状態から遠ざけるという手法を提案している．ここではスケジューリングの有無による結果を示し，この課題の困難さを理解する．1 試行は，パサーがパスをし終えた場合，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超えた場合のいずれかとする．また状態の観測は一定時間間隔ごとに行っている．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，つまり約 0.1 秒ごとに状態を観測すると想定し

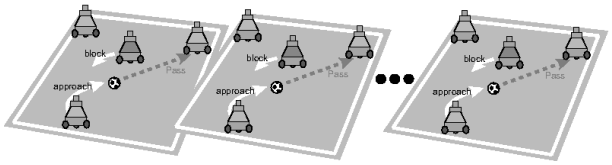
た．また，信頼度を計算するための一定期間 T は以降の実験では 3 ステップ，約 0.3 秒に固定した．

ここではパサーの学習に焦点を当て，インタセプタの行為は設計者が計画をたてた．インタセプタは「右側へのパスをブロックする」方策と「左側へのパスをブロックする」方策を持ち，試行が始まるときにどちらかを選択した後は試行が終わるまでは方策を切り替えない．選択した方策が同じ場合の試行の間では状態遷移確率は変わらないが，方策が異なる試行間では状態遷移確率が変わる．したがって，学習者は相手の方策に応じて学習器を切り替え，状態遷移確率が同じと見なせるような状況に一つの学習器を割り当て，その状況に合った方策を獲得する必要がある．単一の学習器のみではインタセプタの方策の切り替えに対応できないことを実験によって示す．

試行中，レシーバは動かずに静止している．まずインタセプタは右のレシーバへのパスを阻止するための行為を 250 試行続けて取る（**Fig.5**(a)）．次に左のレシーバへのパスを阻止するための行為を 250 試行続けて取る（Fig.5(b)）．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 方策に従う．ある程度各学習モジュールの学習が進んだと思われる 500 試行後，右か左を阻止する行為をランダムに取り，学習者は学習モジュールの信頼度を基に学習すべきモジュールを選択する．



(a) Training pattern 1 : Interceptor is blocking the right side



(b) Training pattern 2 : Interceptor is blocking the left side

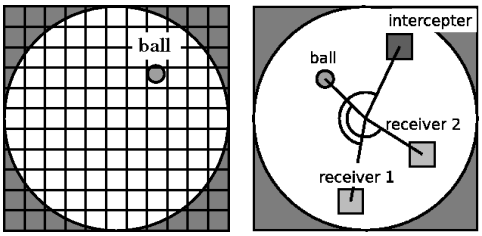
Fig. 5 Training scheduling : One fixed training pattern is shown to learner within a period, the other training patten is shown afther that within the same period, then random pattern is given to the learner

4.1 状態・行動空間設定

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から構成する．**Fig.6**(a),(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 個に等間隔に離散化した．ボールは一つ，エージェントは相手 1 体，味方 2 体の計 3 体なので，$|\mathcal{S}| = 11^2 \times 8 \times 8 \times 8 = 61,952$ となる．

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軸周りの回転運動 (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.7**）．これより $|\mathcal{A}| = 3^3 = 27$ となる．

パサーがボールまでたどり着いたときに，レシーバが正面に見え，かつインタセプタがボールと正面のレシーバの間にいない場合に，学習者（パサー）に対する報酬は 1 とし，それ以外は 0 とした（**Fig.8**）．



(a) Position (b) Angle

Fig. 6 State variables

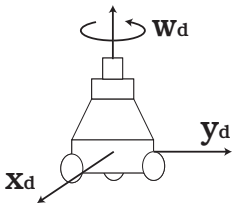


Fig. 7 Action space

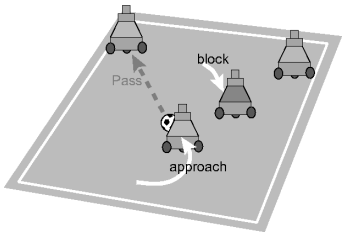


Fig. 8 The situation on which the learner receives a reward

4.2 実験結果

学習により獲得された行動の様子を **Fig.9**，**10** に示す．Fig.9 はインタセプタが学習者（パサー）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（1）〜（3）．その結果右側の味方へのパスを成功させている（4）．**Fig.10** 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．さきほどの左側をブロックした行動と同様に，相手が右側をブロックしているので，学習者はボールに対して右側から回り込んでいる（1）〜（3）．そして左側にいる味方へパスを行なっている（4）．また，**Fig.11** 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，式（5）で計算される各学習モジュールの信頼度の変化を示している．Fig.11(a) を見るとタスク開始の 2 秒から 10 秒にかけて学習モジュール 1 の信頼度が学習モジュール 2 の信頼度よりも大きな値となり，学習モジュール 1 を用いて行動計画を行っている．その結果，Fig.9 で見たようにパス行動を成功させている．同様にして Fig.11(b) においては学習モジュール 2 の信頼度が学習モジュール 1 の信頼度を上回り，学習者は学習モジュール 2 を用いて行動計画を行い，パス行動を成功させて

いる．以上のことから学習モジュール 1 がインタセプタが左に，学習モジュール 2 がインタセプタが右にブロックしてきた状況に対応し，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分かる．Fig.11 において 10 秒当たりから信頼度が落ちているのは，10 秒当たりにアプローチ行動を完了し，ボールを蹴ったため，問題設定から行動学習は終了し，その後は学習中に経験していない状態遷移が行われたためである．形が両試行で異なるのは，学習時の行動選択が ϵ グリーディである程度ランダムな行動を取り，両学習器で経験する時系列が同じでないため，状態遷移モデルも多少ばらつきがあり，そのため信頼度も異なった値を返すためである．

また複数学習モジュールを用いたシステムと単一学習モジュールを用いたシステムについて，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.12**）．スケジューリングなしの学習ではインタセプタが制御方策を毎回交互に替えるという状況で学習する．

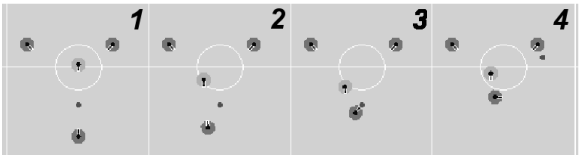


Fig. 9 An acquired behavior against left-block policy of interceptor

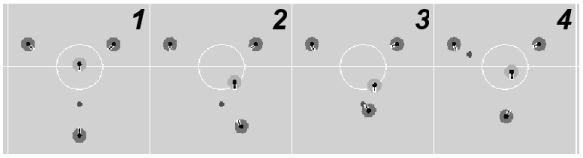
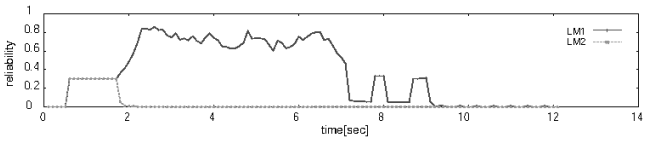
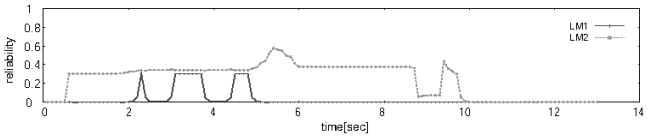


Fig. 10 An acquired behavior against right-block policy of interceptor



(a) The opponent is blocking the left side



(b) The opponent is blocking the right side

Fig. 11 The sequence of the reliabilities of the learning modules

インタセプタが常に左のレシーバへのパスを阻止するための行為を取るようにスケジューリングした 250 試行までのタスク成功率を見ると，モジュール型および単一モジュール型の大差はなく成功率が上昇している．250 試行から 500 試行にかけては単一学習モジュール型は常に失敗を続ける．これは単一の学習モジュールで学習を行い 250 試行までにある状況に特化した方策を持ってしまうため，250 試行以降のインタセプタの方策が変化するという状況の変化に対応できていないと考えられる．それに対して複数学習モジュールの成功率は，状況を区別し新たな状況に対して行動学習し始めるため，それまでに獲得した方策に影響を受けることなく学習を進め，300 試行当たりから成功率は上昇に転じている．500 試行以降は，複数学習モジュール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 80%の成功率となっており，単一学習モジュールを上回っているのがわかる．

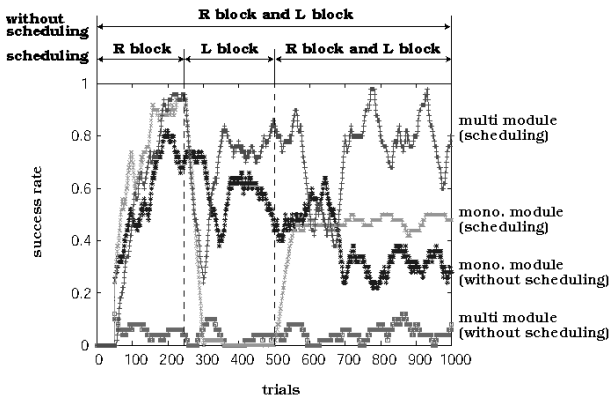


Fig. 12 Curves of success rate

複数学習モジュール手法で，学習のスケジューリングを行う場合と行わない場合を比較すると，後者はかなり低いパフォーマンスを示していることが Fig.12 から分かる．これは学習初期から複数学習モジュールの手法を用いると，各学習モジュールはタスクを遂行するのに適したものにならないということを示している．学習初期においては各学習モジュールの予測能力は低く，それを元に計算される信頼度は正しくなく，それを基に学習モジュールのモデルを構築しても，各学習モジュールは状況を反映した学習モジュールに分かれないと考えられる．

単一学習モジュールにおいて，学習のスケジューリングの有無の影響を見ると，スケジューリングありの場合は，最初に行った固定トレーニングの影響をひきずる結果となったが，スケジューリングなしだと，学習モジュールは一つの状況にひきずられるということはなくなっている．しかし試行が進むに従って成功率が下がっているので，本タスクで与えたような異なる状況を，単一の学習モジュールで学習するのが困難であることを示しており，複数学習モジュールを用いたほうがパフォーマンスが高いことが分かる．

スケジュールなしの条件で，単一学習モジュールが複数学習モジュール手法よりも良い性能を示しているのは，単一モジュールでは二つの状況を混同しても，どちらの状況でも同じ行動をとることが合目的である場合があり成功につながっているが，複数学習モジュール手法では先に述べたように誤った信頼度に基づいて各モジュールを更新しているため首尾一貫した状態遷移を各モジュールで獲得できてないためだと考えられる．

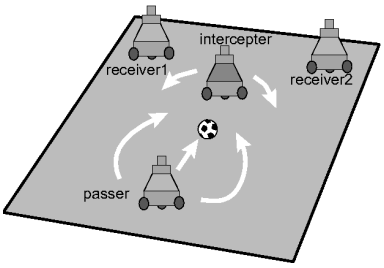


Fig. 13 Macro actions

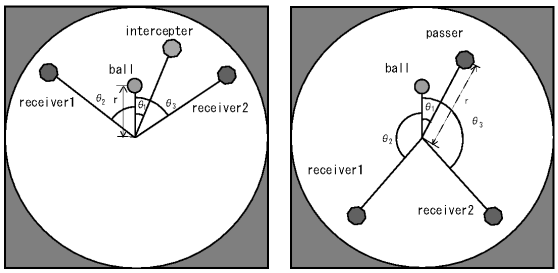
5. マクロ行動をもちいた競合行動の同時学習

前章の実験結果から状況判断のために必要なパラメータ（ここでは状態遷移モデルのパラメータ）空間が大きすぎるため，これらのパラメータの同定にかかる時間が長くかかる．そのため同定する前に他者の方策の変化が起きたので学習が収束せず，パフォーマンスが悪い方策しか獲得できないと推測される．

そこで，ここでは前進後退などの低次の行動ではなく，ボールに近付く，右に回り込むなどのマクロ行動を導入することで，同定しなくてはならないパラメータ数を減らし，結果的に競合行動の同時学習が可能な場合があることを示す．ここでは 2 章で記述した環境において，パサーおよびインタセプタが同時にそれぞれのタスクへの行動を学習する．つまり，パサーはインタセプタの動きに応じてパスをする行動を切り替え，インタセプタはパサーの動きに応じて左右のどちらのパスコースを塞ぐかの行動を学習する．

5.1 マクロ行動と状態空間

パサーとインタセプタのマクロ行動を **Fig.13** に示す．インタセプタのマクロ行動はパサーの右側のレシーバへのパスルートを塞ぐ行動か左側への行動かの 2 種類である．一方でパサーのマクロ行動はボールの周りを左側に回るか右側に回るか，もしくはキックできるように近付くかの 3 種類である．したがって行動数は 4.1 節で説明した 27 から 2 または 3 に減少する．**Fig.14** にパサーとインタセプタの状態空間を示す．パサーの状態空間は前方視覚の画像上でのボールの高さ y と全方位視覚の画像上でのボールの位置とインタセプタの位置の角度，ボールとレシーバの角度（二人のレシーバに対してそれぞれ）で構成される．どのマクロ行動をとっているときもボールが真正面に見えるように姿勢を制御する．このため学習者からボールの方向に関する情報を状態変数に入れる必要がなく，またマクロ行動を利用することで状態を粗く分割することができるので，状態数は 4.1 節で示した 61,952 から 3,773 に減少する．インタセプタの状態空間は前方カメラの画像上でのパサーの高さ y，全方位視覚の画像上でのボールの位置とインタセプタの位置の角度，ボールとレシーバの角度（二人のレシーバにたいしてそれぞれ）で構成され，状態数は 2,695 になる．



(a) Passer (b) Interceptor

Fig. 14 State variables

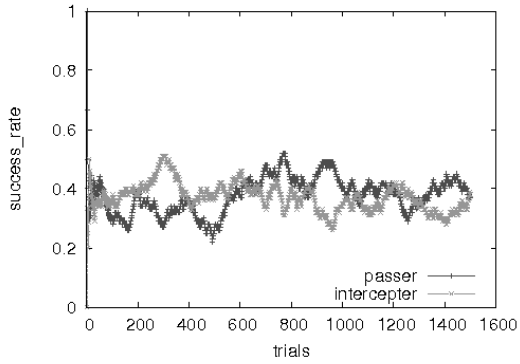


Fig. 15 Sequence of success rates during simultaneous learning

5.2 実験結果

まずシミュレーション上でパサーとインタセプタの同時学習の様子を観察した．両者とも始めから同時に学習し，学習中の行動のスケジュールを行わずに 1,500 試行間学習させた．またこの実験ではあらかじめ学習モジュールを二つ用意し，式（5）および（6）に基づく信頼度によって行動選択および更新のモジュールを切り替えた．学習中のパサー及びインタセプタのタスク成功率を **Fig.15** に示す．両者とも失敗するケースがあるためこの図の両者の成功率を足しても 100%にはならない．600 試行当たりまでインタセプタのほうが先に学習が進み，成功率が高く，しだいにパサーがインタセプタの行動に対応して学習し，1,000 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．

両者が相手の行動選択の状況に応じた合目的的な行動を獲得していることを示すため，パサーかインタセプタの制御方策をあるひとつに固定し，他者がそれに応じた行動を示せるかどうかを調べた．**Fig.16** にその結果を示す．

両者とも二つの学習モジュールを持ち，他者の制御方策によって引き起こされる状況に対応できるようになっている．LM とその後に続く数字はそれぞれ学習モジュールとそのインデックスを示す．例えばインタセプタが LM0 学習モジュールのみを使い，パサーは LM0 および LM1 の両方の学習器を使った場合，パサーの成功率が 59%，インタセプタの成功率が 23%，両方とも失敗する確率が 18%となっている．この図から明らかにパサーにとってもインタセプタにとっても複数学習モジュールを持ちそれらを切り替えるほうが成功率が高いことが分かる．

実ロボットに対しても同じ構成で実験を行った．**Fig.17** にその実験の様子を示す．まず，インタセプタが左側のパス経路を塞ぐように動き，パサーは右側にパスを出すようにボールに近付く．パサーの行動を見て，そこでインタセプタは逆方向を防ぐように動き出す．しかしパサーが先に右側のレシーバにパ

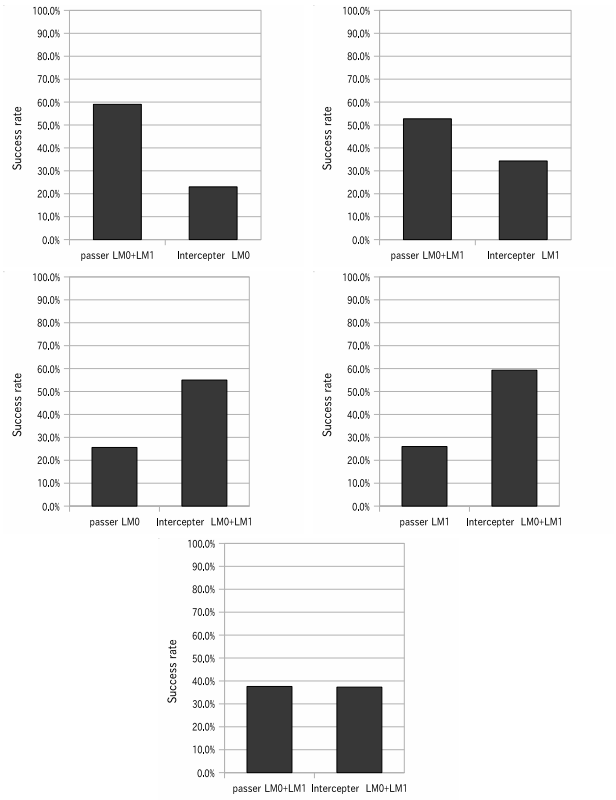


Fig. 16 Success rates for passer and receiver in different cases

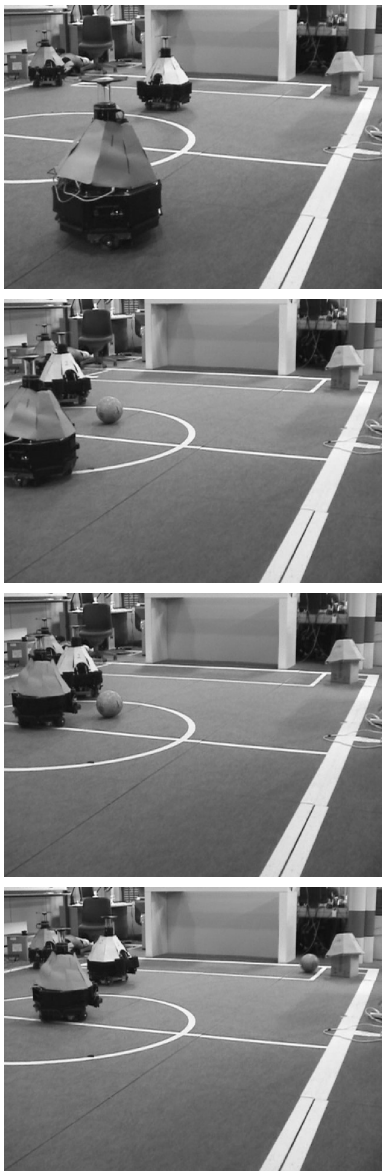


Fig. 17 A sequence of a behavior of passing a ball to the right receiver while the interceptor blocks the left side

スを通して成功させている．

6. 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獲得を実現するため，まずマルチモジュール型学習機構を適用した．他者の制御方策変動に対して，学習者から見た環境の状態遷移確率が大きく変化するような状況においても，学習のスケジューリングを導入することにより，合目的的な行動獲得が可能であることを示した．また探索空間を小さくすることで競合行動の同時学習が可能である場合があることを示した．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モジュールに割り当てる．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，相手の制御方策の変化に対し学習モジュールの切り替えがうまくいき，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．

ここで，相手の方策変動が観測できる場合，それに伴う状況変化を含めた単一学習器での学習可能性を検討する．この場合，状態遷移が自己の行動選択の結果なのか，他者の方策変動によるものであるのか区別しなければならない．また，区別された自己の行動選択に対して適切な評価を与える必要もある．さらにより重要なことは，他者の方策変動に対する行動学習への悪影響を抑える仕組みを組み込まなければならない．この影響を考慮しない場合，相手の方策変動に影響を受ける状況では自己の行動選択に対する評価が正しく行われず，合目的的な行動が獲得されない可能性がある．モジュール型学習機構の場合は，状況に応じて学習器を切り替え，状況の切り替えはどの学習器も責任を負わないことで，この問題を回避できる．

本稿では，マルチエージェントの競合行動によって引き起こされる，学習者から見た環境の動的変化の問題に対し，マルチモジュール型学習機構を用いた動的環境への適応という観点から，相手の出方に応じた適切な行動獲得を実現した．今後は，Bagging などの分類器生成の手法を組み合わせることで，性能向上や適用範囲の拡大を目指す．スケジューリングの自動設定 [13] やマクロ行動の自己組織化 [25–27] などの研究成果を取り入れ，よりロバストな行動学習の可能性を探る．また，むしろ相手の出方を誘発する行動を自ら積極的に行い，自身にとって有利な状況をつくり出す行為（フェイント）を獲得する手法について

研究を進める予定である．

参 考 文 献

[1] Jonathan H. Connell and Sridhar Mahadevan. *ROBOT LEARNING*. Kluwer Academic Publishers, 1993.

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

[3] Morimoto J. and Doya K. Hierarchical reinforcement learning of low-dimensional subgoals and high-dimensional trajectories. In *The 5th International Conference on Neural Information Processing*, Vol. 2, pp. 850–853, 1998.

[4] Keishiro Tabe, Kenji Suzuki, Pitoyo Hartono, and Shuji Hashimoto. Survival strategy learning for autonomous mobile robot. In *Proceedings of of the 2001 IEEE-RAS International Conference on Humanoid Robots (Humanoids2001)*, Japan, Nov 2001.

[5] Katsunari Shibata and Masaru Iida. Acquisition of box pushing by direct-vision-based reinforcement learning. In *Proceedings of SICE Annual Conference in Fukui*, Vol. CD-ROM, pp. 1378–1383, Aug 2003.

[6] J. Morimoto, G. Zeglin, and C. G. Atkeson. Minimax differential dynamic programming: Application to a biped walking robot. In *Proceedings of SICE Annual Conference in Fukui*, Vol. CD-ROM, pp. 584–586, Aug 2003.

[7] Cody Kwok and Dieter Fox. Reinforcement learning for sensing strategies. In *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. CD–ROM, Sep 2004.

[8] Nate Kohl and Peter Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *In Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 2619–2624, May 2004.

[9] 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.

[10] Eiji Uchibe and Kenji Doya. Reinforcement learning with multiple heterogeneous modules: A framework for developmental robot learning. In *Proceedings of 2005 4th IEEE International Conference on Development and Learning*, pp. 87–92, 2005.

[11] Maja J Mataric. Reinforcement learning in the multi robot domain. *Autonomous Robots*, Vol. 4, No. 1, pp. 77–83, 1997.

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

[13] 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, Oct 2002.

[14] 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.

[15] R. Jacobs, M. Jordan, Nowlan S, and G. Hinton. Adaptive mixture of local experts. *Neural Computation*, Vol. 3, pp. 79–87, 1991.

[16] Satinder Pal Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, Vol. 8, pp. 323–339, 1992.

[17] Satinder P. Singh. The effeicient learnig of multiple task sequences. In *Neural Information Processing Systems 4*, pp. 251–258, 1992.

[18] Jun Tani and Stefano Nolfi. Self-organization of modules and their hierarchy in robot learning problems: A dynamical systems approach. Technical report, Technical Report: SCSL-TR-97-008, 1997.

[19] J. Tani and S. Nolfi. Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems. *Neural Networks*, Vol. 12, No. 7-8, pp. 1131–1141, 1999.

[20] Klaus-Robert Muller, Kohlmorgen Jens, and Pawelzik Klaus. Analysis of switching dynamics with competing neural networks. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, Vol. E78-A, No. 10, pp. 1306–1315, Oct 1995.

[21] P. Hartono and S. Hashimoto. Temperature switching in neural network ensemble. *Journal of Signal Processing*, Vol. 4, No. 5, pp. 395–402, 2000.

[22] 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, No. 5, pp. 551–556, 2001.

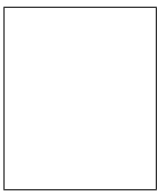
[23] Masahiko Haruno, Daniel M. Wolpert, and Mitsuo Kawato. Mosaic model for sensorimotor learning and control. *Neural Computation*, Vol. 13, pp. 2201–2220, 2001.

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

[25] Yasutake Takahashi, Koichi Hikita, and Minoru Asada. Incremental purposive behavior acquisition based on self-interpretation of instructions by coach. In *Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 686–693, Oct 2003.

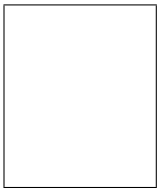
[26] Tomoki Nishi, Yasutake Takahashi, and Minoru Asada. Incremental purposive behavior acquisition based on modular learning system. In Tamio Arai, Rolf Pfeifer, Tucker Balch, and Hiroshi Yokoi, editors, *Intelligent Autonomous Systems 9*, pp. 702–712. IOS Press, March 2006. ISBN 1-58603-595-9.

[27] Tomoki Nishi, Yasutake Takahashi, and Minoru Asada. Incremental behavior acquisition based on reliability of obser ved behavior recognition. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 70–75, Oct 2007.



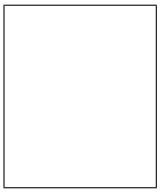
高橋 泰岳（Yasutake Takahashi）

1994 年大阪大学大学院工学研究科博士前期課程修了．2000 年同大学博士後期課程中退，同年同大学大学院工学研究科助手．現在大阪大学大学院工学研究科知能・機能創成工学専攻助教．この間 2006 年 6 月より 2007 年 9 月までドイツ Fraunhofer IAIS 客員研究員．ロボカップ中型機リーグや知能ロボットの行動獲得に関する研究に従事．人工知能学会，知能情報ファジィ学会などの会員．（日本ロボット学会正会員）



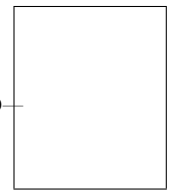
枝澤 一寛（Kazuhiro Edazawa）

2002 年大阪大学工学部応用理工学科卒業．2004 年大阪大学大学院工学研究科知能・機能創成工学専攻修了．2004 年三菱電機株式会社入社．現在，先端技術総合研究所で，物理セキュリティシステム開発業務に携わる．



野間 健太郎（Noma Kentaro）

2005 年大阪大学大学工学部応用理工学科卒業．2007 年大阪大学大学院知能・機能創成工学専攻修了．2007 年富士フイルム入社．現在，R & D 統括本部機器システム開発センターで，医療機器開発業務に携わる．



浅田 稔（Minoru Asada**）**

1982 年大阪大学大学院基礎工学研究科後期課程修了．1995 年大阪大学工学部教授．1997 年大阪大学大学院工学研究科知能・機能創成工学専攻教授となり現在に至る．2005 年より JST ERATO 浅田共創知能システムプロジェクト研究総括．認知発達ロボティクスの研究に従事．本学会論文賞（1996），文部科学大臣賞（2001）など受賞多数．