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 (大阪大)　高橋 泰岳 (大阪大)　浅田 稔 (大阪大)

Multi-Module Learning System for Behavior Acquisition in Multi-Agent Environment

*Kazuhiro EDAZAWA (Osaka Univ.), Yasutake TAKAHASHI (Osaka Univ.), and Minoru ASADA (Osaka Univ.)

Abstract— The conventional reinforcement learning approaches have difficulties in handling the policy alternation of the opponents because it may cause dynamic changes of state transition probabilities of which stability is necessary for the learning to converge. If we can assign multiple learning modules to different situations in which the each module can regard the state transition probabilities as consistent, then the system would provide reasonable performance. This paper presents a method of multi-module reinforcement learning in a multiagent environment, by which the learning agent can adapt its behaviors to the situations as results of the other agent's behaviors. We show a preliminary result applied to a simple soccer situation.

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

1. はじめに

これまで強化学習を用いた行動獲得の研究が多くなされてきた [1)2)]．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．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環境下においての合目的的行動の獲得は従来手法では難しい．

銅谷ら [3)] 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．そこで本論文ではこの基本的な考えを，マルチエージェント環境における行動獲得に応用する．つまり学習者に複数の学習モジュールを持たせ，各モジュールを状態遷移確率が一定とみなせる状況に個別に割り当てることにより，他のエージェントの行為が変化するような環境下で合目的的な行動が獲得できることを示す．簡単なサッカーのタスクを通じて提案手法の有効性をシミュレーションで示す．

2. 提案手法

提案手法を Fig.1 に示す．点線で囲まれた各モジュールは予測器 (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，計画器は状態遷移確率モデルに基づいて強化学習の手法で行動価値関数を推定する．システムは状態遷移を最も良く予測しているモジュールを選択し，その状況にあった行動をとる．

2·1 予測器

各モジュールは予測モデルを持ち，ある状態 s，行動 a，次状態 s' となる確率 $\hat{\mathcal{P}}^a_{ss'}$ を経験をもとに推定する．またシステムは状態予測モデルだけでなく報酬モデル $\hat{\mathcal{R}}^a_{ss'}$ も持つ．

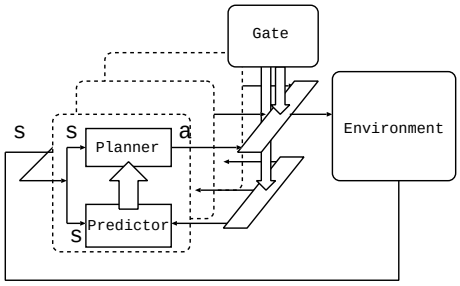


Fig.1 A multi-module learning system

2·2 学習器

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 $Q(s,a)$ は以下のようにして与えられる．

$$Q(s,a) = \sum_{s'} \hat{\mathcal{P}}^a_{ss'} \left[\hat{\mathcal{R}}^a_{ss'} + \gamma \max_{a'} Q(s',a') \right] , \quad (1)$$

ここで γ は減衰係数を表す．

2·3 ゲート信号

現在の状況を最も良く予測しているモジュールはその状況に対して最も良い政策を持つと考えられる．なぜならそのモジュールの計画器は今の状況に最も適合しているモデルを基に政策を計算しているからである．最も良く予測しているモジュールを選択するために，ここではゲート信号を導入する．ゲート信号はそのモジュールの予測器が状況を良く予測していると大きな値になり，そうでないと小さな値となる．モジュール i のゲート信号 g_i は以下のように計算される．

$$g_i = \prod_{t=-T}^{1} \frac{e^{\lambda p_i^t}}{\sum_j e^{\lambda p_j^t}}$$

ここで p_i^t はモジュール i のモデルによる時刻 $t-1$ の状態から時刻 t の状態への状態遷移確率の予測値であり，λ は定数である．

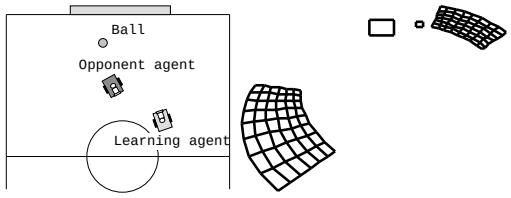
3. シミュレータによる実験

提案手法の有効性を検証するために，学習者が敵を避けながらボールまで到達するタスクを予備実験としてシミュレーションで行った．

3·1 実験環境

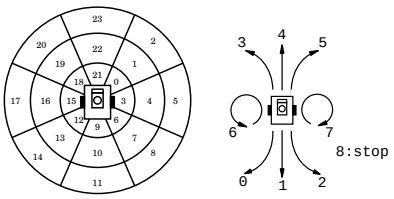
ロボットは RoboCup に実際に出場しているものを想定する．このロボットは全方位カメラシステムを持っており，33ms 毎にボールと敵を検知することができる．Fig.2(a) に学習中の状況を示しており，Fig.2(b) はロボットの全方位カメラの画像をシミュレーションした様子である．大小の長方形はそれぞれ敵とボールを示している．状態空間はボールと敵の画像上の位置を角度と距離方向で離散化し構成する (Fig.3(a))．駆動機構は PWS(Power Wheeled Steering) システムを用い，行動空間は二つの車輪のトルクの値を離散化し構築する (Fig.3(b))．

敵は “stop”, “move left”, “move right” というような行動を一定間隔毎に切り変える．また学習者は敵のこういった行動毎のモジュールを割り当てられている．学習中はこれらのモデルを構築するために，ランダムに動き回ってボールと敵の画像情報を集める．



(a) Top view (b) Simulated image

Fig.2 Simulation Environment



(a) State space (b) Action space

Fig.3 State-action space

3·2 実験結果

本手法を用いてたロボットと，単一のモジュールしか持たないロボットとを比較した．Table.1 は学習後のタスク達成率をそれぞれ示している．ここでタスクの成功とは敵がランダムに動いている時に，敵を避けてボールまで辿り着くということを指す．成功率は 100 回の試行におけるタスクの成功する割合である．結果を見ると単一のモジュールしか持たないものよりも複数のモジュールを持った学習者の方が高い成功率をしている．Fig.4 は学習者は学習済みの政策を用いて行動

Table 1 Success rates between multi-module system and one-module one

system	success rate
multi-module	61 %
one-module	50 %

し，敵はランダムに行動したときのゲート信号を表している．Fig.4 を見ると最初と最後で敵の行動の予測を誤っているように見えるが，学習者はタスクを成功している．これは，たとえ学習者が敵の行動の予測を誤ったとしても，学習者の政策が他のエージェントの行動と関係のない状況であればそれは問題ない，ということを意味している．

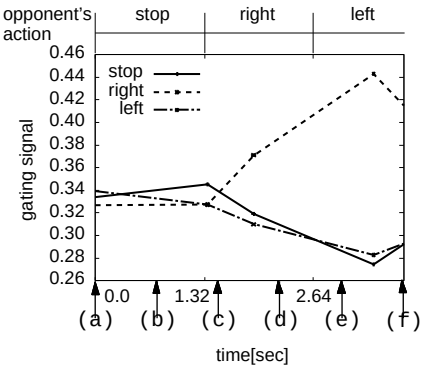


Fig.4 A sequence of gating signal while the agent executes its learned policy

4. まとめ

本研究では，他のエージェントの政策の変化による状況の変化に対して，複数のモジュールを割り当てることによりその状況に適した行動を学習させ，その有効性をサッカーのタスクを用いてシミュレーションで示した．今回はモジュールの数を固定にして状況に応じて割り当てを行ったが，将来は学習者自身がモジュールを獲得するようにする予定である．

謝辞

本研究は科学技術振興事業団の戦略的基礎研究推進事業「脳を創る」中村プロジェクトの援助を受けた．

参考文献

1) 細田耕内部英治. 複数の学習するロボットの存在する環境における協調行動獲得のための状態空間の構成. 日本ロボット学会誌, Vol. 20, No. 3, pp. 281–289, 2002.
2) 高橋, 浅田. 複数の学習器の階層的構築による行動獲得. 日本ロボット学会誌, Vol. 18, No. 7, pp. 1040–1046, 2000.
3) 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.