非同期政策更新に基づくマルチエージェント同時強化学習による協調行動の獲得

○池上渉一 (大阪大学, 現：ソニー (株)) 浅田稔（大阪大学) 細田耕（大阪大学）

Simultaneous learning for cooperative behavior acquistion in multiagent environment based on asyncronous policy renewal

*Shoichi Ikenoue (Osaka Univ.,currently with SONY),
Minoru Asada(Osaka Univ.), Koh Hosoda(Osaka Univ.)

Abstract—This paper presents a method for simultaneous learning in multiagent environment to emerge the cooperative behaviors. Each agent has one policy and one action value function: the former is for action execution based on the action value function updated in the previous stage, and the latter is for learning based on the episodes experienced by the current policy. This makes all agents behave based on the fixed policies, by which the non-Markovian problem can be avoided except for the update periods that depends on the learning progress of each agent. In order to avoid the local maxima due to such asynchronous renewal of action value functions, optimistic action values are given as initial ones, that helps the exploration process not to be trapped in the local maxima. The experimental results applied to one of the cooperative task in dynamic, multiagent environment, RoboCup, is shown and a discussion is given.

Key Words: multiagent・cooperative behavior・reinforcement learning・simultaneous learning・RoboCup

1. はじめに

複数エージェントの協調行動は，エージェントがタスクを遂行する上で重要である．この協調行動の獲得に自律的に行動を獲得することのできる強化学習 [1] を適用することが近年議論されている [2][3]．しかし，強化学習の適用に際しては，学習環境がマルコフ性を満たしていることが必要とされるが，複数のエージェントが同時に強化学習を行った場合各自が学習中に試行錯誤を繰り返す為，この条件を満たすのは困難である．

これに対し，Asada et al.[4] は中央集権的に学習者をスケジューリングすることで，先の問題をクリアし，強化学習を用いて複数エージェントでの協調行動の獲得に成功している．しかしながら中央集権的な学習者の切替えは，エージェントの自律性の観点から出来る限り避けたい．

そこで本研究では，エージェントに自分の学習がいったん収束するまで自分の行動政策を固定させ，それをエージェント同士が互いに行うことで，他者の行動政策を一定期間固定にし，さらに，行動政策の更新方法と学習環境の探査方法を工夫することで，複数のエージェントが同時に強化学習を行うことを可能にする枠組みを提案する．

提案手法を近年人工知能と知能ロボットの分野での標準問題として提唱されているロボカップ [5] 中型機リーグのレギュレーションのもと，複数エージェントの協調行動として，パス&シュート行動を 2 台のロボットに強化学習によって獲得させる．シミュレーションによる実験を行い，実機での検証でその有効性を示す．

2. 非同期政策更新に基づく同時学習

学習アルゴリズムは広く利用されている Q 学習 [6] を用いる．この時，学習エージェントは，一つ前の学習エピソードで得られた行動価値関数 Q^{n-1} に基づいた行動政策 π^n を用いて，学習を行うことで，学習結果を即時に行動政策に反映させるのではなく，一定期間経験を蓄積させ，ある程度収束した時刻で，それを新たな行動政策として更新させる．これにより，環境中に存在する全てのエージェントの行動政策は一定期間固定となる為，他者が行動政策を更新するまでの間，個々の学習者が観測する学習環境はマルコフ性を満たす．すなわち，マルコフ性は環境に存在するエージェントの中で，一番最初に政策を更新した者に対して保証される．

このことから，同じエージェントが続けて政策を更新しつづけることで，学習の進度に差が出るのを防ぐ必要が生じる．また，常に固定政策で学習を行うことから探査を網羅的にする必要もある．これらのことから，学習の最初に行動価値関数を報酬値以上の値に初期化する手法を導入する．これは，現在の行動政策で選択される行動が現在の価値に見合わない場合に，価値が下がっていくことを利用して行動政策を遷移させ探査を網羅的に行う為であり，ペシミィスティック (Fig.1) な行動価値関数の初期化とは逆の方向から，最適行動価値関数を近似していく方法として，オプティミィスティックな行動価値関数の初期化という名前で知られている [1]．さらにこの時，探査時の行動政策を完全に固定する為に，行動価値関数を予め不均一にしておく必要がある．これによって，行動価値が等価な時に発生するランダム性を排除することができる．これらのことから提案手法において，行動価値関数は，Fig.1(右参照) に示されるように，オプティミィスティックかつ不均一 (1.0〜1.0001) に初期化される．

また Q^n の収束判定を，エージェント自らの入力のみで判断させることで，純粋に各自の入力のみから学

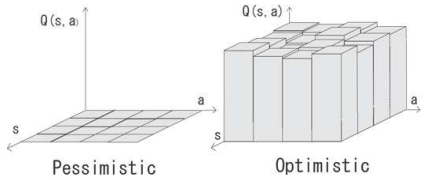

Fig.1 Initialization of action-value function

習を行うことができる．具体的には，

$$\sigma_Q = \sum_s \max_{a \in A} [Q^n(s,a)] \{\forall s \in S\} \qquad (1)$$

と定義される σ_Q を各試行毎に計算し，その推移の傾きを最小二乗法を用いて求め，閾値 θ_{σ_Q} を下回るかどうかで Q^n の収束を判断させる．

以下に 1 ステップの学習アルゴリズムを示す．

1. エージェントは Q^0 を作成し，オプティミィスティックに初期化する．
2. エージェントは π^1 を作成し，Q^0 に基づいて初期化する．
3. エージェントは環境中で状態 $s_t \in S$ を観測する．
4. 現在の状態 s に対し，行動政策 π^n に従って行動 $a_t \in A$ を選択する．

$$\pi^n(s) = arg \max_{a \in A} [Q^{n-1}(s,a)] \{\forall s \in S\} \qquad (2)$$

5. 行動 a_t によって状態が $s_{t+1} \in S$ に遷移する．
6. 環境から報酬 r_{t+1} が与えられ，

$$Q^n(s_t,a_t) \leftarrow Q^n(s_t,a_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - Q^n(s_t,a_t)] \qquad (3)$$

$$V(s_t) = \max_a Q^n(s_{t+1},a) \qquad (4)$$

の様に行動価値観数 Q^n を更新していくことで．最適行動価値関数 Q^* を近似していく．

7. Q^n が収束していたら，それに基づき $\pi^{n+1}(s)$ を更新する．

$$\pi^{n+1}(s) \leftarrow arg \max_{a \in A} Q^n(s,a) \qquad (5)$$

8. 3 に戻る．

3. 実験設定

提案手法を近年人工知能と知能ロボットの分野で標準問題として提唱されているロボカップ [5] に適用する．協調行動の例として，2 台のロボットによるパス&シュート行動の獲得を目的とし，シミュレーションにより実験を行う．

ロボカップの中型機リーグのレギュレーションに従い．縦 8[m]，横 4[m] の壁で囲まれ，両端にゴールが存在するフィールドを用意する．そして Fig.2 に示す様に，学習エージェントとして 2 台の提案手法を組み込んだロボット (Passer，Shooter) をそれぞれ配置し，更に BALL AREA に 10 通りにランダムでボールを配置する．ロボットは全方位視覚センサとキック機構を持っており，移動機構として左右に取り付けられた駆動輪の回転差によってステアリングを切る独立二輪駆動輪操舵方式 (Power Wheeled Steering Method,PWS 方式) を装備している．内部構造を Fig.3 に示す．この際，エージェントはロボットの 33[msec] のサンプリングレートに対応する 1 周期で 1 回の状態を知覚できるが，学習をスムーズに行う為，浅田ら [7] によって提案されている「ある状態で行動選択をした場合，状態遷移が起こるまで同じ行動を継続実行する」手法を用いる．

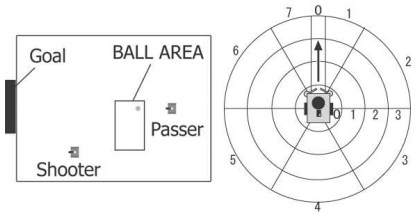

Fig.2 Initial positions in the computer simulator and a state space of the agent

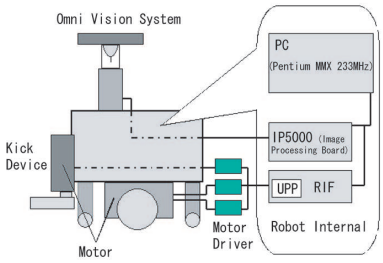

Fig.3 An overview of the robot system

またエージェントの状態空間をロボットの全方位視覚センサから得られる画像を適当に離散化することで構成する．具体的には，ロボットが知覚するボール，敵ゴール，味方について，得られた画像をまず方向ごとに 8 個の領域に分割し，更にそれぞれの領域を，距離ごとに 4 個の領域に分割する (Fig.2：紙面上が進行方向)．この時，前方向の状態を細かく分割するのは，キック機構がロボット前方に装備されていることから，後方よりも前方におけるオブジェクトの位置が重要な為である．そしてこれらを方向 8 状態，距離 4 状態として更にそれぞれを見失った場合の状態を含めて，6 次元，$(8 \times 4 + 1)^3 = 35937$ 状態の状態空間を構成する．この時，それぞれの領域は，進行方向から時計回りに，$0, 1, 2, \cdot\cdot 7$ の数字を割り当て，同じく距離状態は，一番近傍から $0, 1, 2, 3$ と割り当てる．また行動空間は，前進，後退，右回転，左回転の 4 状態とする．キック機構はボールが蹴れる範囲に入ったら自動的に稼働する．

更に報酬関数を

- エージェント 1(Passer) は，パスが成功した時，すなわち状態 $s =$ { ボール方向, ボール距離, 敵ゴール方向, 敵ゴール距離, 味方ロボット方向, 味方ロボット距離 } $= \{0, 0, a, b, c, d\} (a = 0$ or $1, c = 0$ or 7 ,b, d は任意) において前進行動を選択した時，報酬 1 を受けとる．

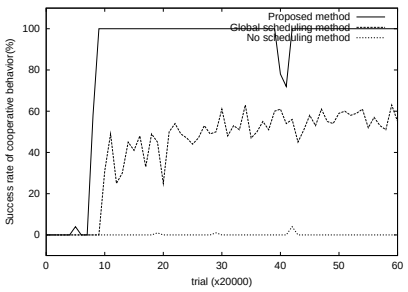

Fig.4 Comparison of scheduling methods

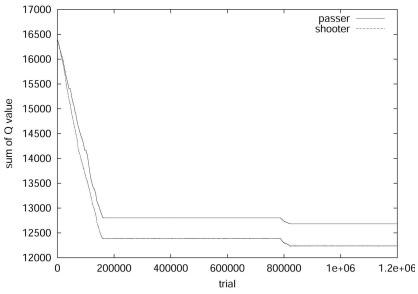

Fig.5 Transition of σ_Q in optimistic initialization

- エージェント 2(Shooter) は，シュートが成功した時，すなわち状態 $s = \{$ ボール方向, ボール距離, 敵ゴール方向, 敵ゴール距離, 味方ロボット方向, 味方ロボット距離 $\} = \{\ 0, 0, a, b, c, d\ \}(a = 0, b, c, d$ は任意) において前進行動を選択した時，報酬 1 を受けとる．
- エージェント同士が衝突した場合，報酬 -1 をエージェント 1 が受けとる．
- それ以外の状態では，行動によらず報酬は 0.

とする．また学習率 $\alpha : 0.2$，減衰率 $\gamma : 0.9$ とする．

4. 実験結果

前節に従い

1. 提案手法 (Proposed method [1.0 〜 1.0001])
2. 中央集権的な学習者のスケジューリング (Global scheduling method [0.0])
3. スケジューリング無し (No scheduling method [0.0])

の 3 つについて協調行動の成功率の推移 (Fig.4) を 1,000 試行ごとにプロットして比較した (それぞれの右の [] の値は行動価値関数の初期値)．提案手法においては，800,000 試行前後で一度成功率が落ちているが，前後で得られている行動を比較した所，報酬に至るまでの状態変化の回数の短縮が見られた．この時の提案手法の θ_{σ_Q} は 0.01 であり，その時の行動政策の 10,000 試行ごとの更新頻度推移と σ_Q の推移を Figs.5 and 6 に示す．

更に，提案手法においては，行動価値関数をオプティミィスティックに初期化することで学習環境の探査を行っていたが，その初期化の影響を調査するために行動価値関数を 0〜0.0001 に初期化して実験を行った．この時，完全に 0 に初期化しないのは，全て 0 にしてしまうと探査が進まない為である．同時に，中央集権的な学習者のスケジューリングを行った場合，スケジューリングを行わなかった場合について，それぞれ行動価値関数を 1.0〜1.0001 に初期化して実験した．この結果，提案手法においてのみ協調行動の成功率の推移に変化が見られ，他の手法に関しては，ほとんど変化は見られなかった．この時の提案手法の行動政策の更新頻度推移及び σ_Q の推移を Figs.7 and 8 に示す．

5. 考察

実験結果より，

- 学習者のスケジューリングを行わない場合，相互に同時に学習を行うことから，個々の学習環境はマルコフ性を満たさない為，行動価値関数の初期値に関わらず協調行動は獲得されない．
- 中央集権的に学習者をスケジューリングした場合，相互に同時に学習を行うことは無く，個々の学習環境はマルコフ性を満たす為，行動価値関数の初期値に関わらず協調行動は獲得される．
- 提案手法においては，学習者のスケジューリングは行っていないが，行動価値関数の初期値と政策更新の閾値を調節する事で，協調行動が獲得される．

の 3 点が確認された．

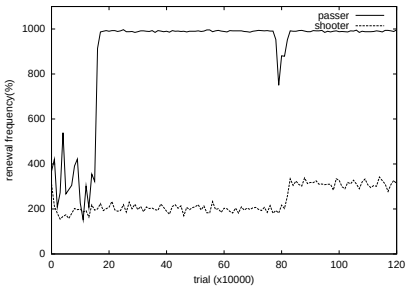

Fig.6 Renewal frequency in optimistic initialization

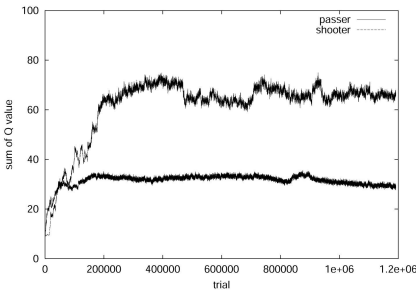

Fig.7 Transition of σ_Q in pessimistic initialization

このうち，3 点目については以下のように解釈できる．提案手法においては，その探査政策は，Q^n の収束性に依存して更新され，かつ更新されるまでの間，Q^{n-1} のグリーディ政策で固定される．このことから，ペシミィスティックに行動価値関数を初期化した場合，エージェントは，学習の初期において非常に狭い範囲の探査しか行わず，σ_Q の変化は多くの場合に微小となるため，政策更新の頻度は高くなる．その結果，後の探査にバイアスがかかることから，非常に局所解に陥る可能性が大きくなる．これに対して，オプティミィス

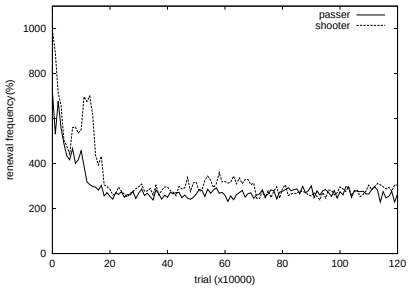

Fig.8 Renewal frequency in pessimistic initialization

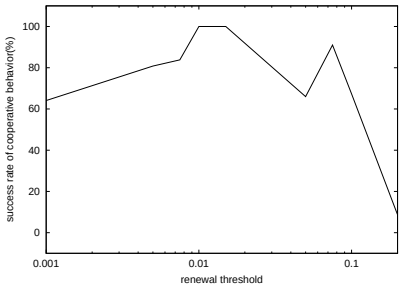

Fig.9 Comparison of threshold value

ティックに行動価値関数を初期化した場合は，エージェントが，学習の初期において網羅的に探査することで，σ_Q の変化が大きくなり，政策更新の頻度は低くなる．その結果，不適当な状態行動対の価値が直ちに減少することから，最適な行動は発見され易くなる．この結果，最終的に行動価値関数は収束し，それに伴い更新頻度は高くなる．Figs.5 and 6 and 7 and 8 は，これらを良く示している．

また，提案手法における各エージェントの収束判定閾値 θ_{σ_Q} は 0.01 と設定されていたが，これについて値を変えて実験した．各閾値での収束結果を見る為に，1,400,000 試行から 1,600,000 試行における協調行動の成功率の平均を縦軸に取り，閾値 θ_{σ_Q} を横軸に取る事で比較した．Fig.9 に見られる通り，更新閾値は学習の収束性に大きく影響する．これは閾値が大きすぎれば，それは更新頻度が非常に高い事を意味する為，スケジューリングをしない事と等価となり，逆に閾値が小さすぎれば，それは更新頻度が非常に低い事を意味する為，長期間同じ政策を試し続ける事から，探査にバイアスがかかり易く，局所解に陥り易くなる為である．これらのことから，閾値の設定は重要な問題である．

最後に，学習された行動政策を実ロボットに実装する事で得られた，協調行動のシーケンスを Fig.10 に示す．シミュレーションと実環境の差異から，成功率は多少減少したが，図の通り協調行動を実現する事ができた．しかし，この差を完全に無くす事は非常に困難であることから，直接実環境で実験を行っていく為にも，今後学習時間の短縮について取り組む必要がある．

6. おわりに

本論文では，マルチエージェント環境において同時学習を可能とする枠組みを提案し，ロボカップ（サッカータスク）におけるパス&シュート行動に適用した．

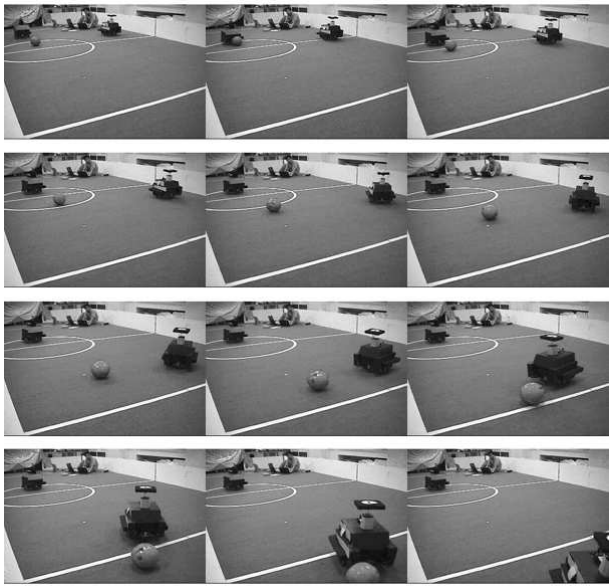

Fig.10 A sequence of cooperative behavior by real robots

主な特徴は，個々のエージェントが自己の学習が収束判定されるまで探査政策を固定し，その間に得られた行動価値関数を次の探査政策として利用する事で，複数のエージェントが同時に学習を行っても，環境がマルコフ的に保たれる様にする事である．これにより，中央集権的な学習者のスケジューリングを行う必要なしに，自己の行動価値関数の収束判定のみに従い，探査政策をそれぞれ非同期に更新しあう事で，協調行動が獲得される．但しこの際，学習の収束性は収束判定閾値に大きく依存する為，今後，この閾値を決定する系統的な方法を開発していく必要がある．

参考文献

1) 三上貞芳, 皆川雅章訳:”強化学習”，森北出版株式会社，2000(Richard S.Sutton and Andrew G.Barto: “Reinforcement learning:An Introduction”, MIT Press/Bradford Books, March 1998).
2) Peter Stone and Richard S.Sutton:“Scaling Reinforcement Learning toward RoboCup Soccer”, 18th International Conference on Machine Learning，2001.
3) Yasuo Nagayuki,Shin Ishii and Kenji Doya: “Multi-Agent Reinforcement Learning:An Approach Based on the Other Agent’s Internal Model”, Fourth International Conference on MultiAgent System(ICMAS) Los Alamitos:IEEE Computer Soceiety, pp.215-221,2000.
4) M.Asada,E.Uchibe and K.Hosoda: “Cooperative Behavior Acquistion for Mobile Robots in Dynamically Changing Real World via Vision-Based Reinforcement Learning and Development”, Artificial Intelligence, Vol.110,pp.275-292,1999.
5) 松原仁, 浅田稔, 北野宏明: “ロボカップの歴史と 2002 年への展望”. 日本ロボット学会誌,Vol.20,No.1,pp.2-6,2002.
6) C.J.C.H.,Watkins and Dayan P: ”Technical note:Q-learning”, Machine Learning, Vol.8,pp.279-292,1992.
7) 浅田稔，野田彰一，俵積田健，and 細田耕: “視覚に基づく強化学習によるロボットの行動獲得”, 日本ロボット学会誌，Vol.13,No.1,pp.68-74,1995.