他者の状態価値の推定に基づく協調・競合行動の獲得

Cooperative/Competitive Behavior Acquisition Based on State Value Estimation of Others

野間 健太郎[1]，**高橋泰岳**[1]，**浅田 稔**[1,2]

Kentaro NOMA[1], **Yasutake TAKAHASHI**[1], **Minoru ASADA**[1,2]

[1] 阪大学大学院工学研究科，[2]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,** [2]**JST ERATO Asada Synergistic Intelligence Project**

{kentaro.noma,yasutake,asada}@ams.eng.osaka-u.ac.jp

Abstract

The existing reinforcement learning approaches have been suffering from the curse of dimension problem when they are applied to multiagent dynamic environments. One of the typical examples is a case of RoboCup competitions since other agents and their behaviors easily cause state and action space explosion. The keys for learning to acquire cooperative/competitive behaviors in such an environment are as follows:

- a two-layer hierarchical system with multi learning modules is adopted to reduce the size of the sensor and action spaces. The state space of the top layer consists of the state values indicating how close to the goals of the individual modules at the lower level, and the macro actions are used to reduce the size of the physical action space, and further,
- to what extent the other agent task has been achieved is estimated by observation and used as a state value in the top layer state space to accelerate the cooperative/competitive behavior learning.

This paper presents a method of modular learning in a multiagent environment, by which the learning agent can acquire cooperative behaviors with its team mates and competitive ones against its opponents. The method is applied to 4 on 5 passing task, and the learning agent successfully obtained the desired behaviors.

1 はじめに

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，協調・競合行動の獲得を行う研究が多くなされている[1, 2, 3, 4, 5]. マルチエージェント環境で強化学習を適用する際の問題点として，自身や対象物の記述だけでなく，複数の他者との関係の記述も必要なため，考慮しなければならい情報が多くなり，探索空間が莫大になるため現実時間で学習するのが困難である．

Shivaram et al.[3]は，ハーフコートのサッカーフィールドで 4 対 5 でパスを行いシュートを決めるタスクで，味方の学習情報を共有することで，学習効率が上がることを示した．しかし，センサレベルの状態変数を使って状況判断をしているため，先に述べたように探索空間が大きく，学習時間が非常に長い．Stefan et al.[1]はマクロ行動を導入することにより，2 台のロボットが協調行動の獲得を実時間で実現している．マクロ行動とは設計者によってあらかじめ決められた行動のことで，モータレベルの行動を学習する必要がないので，効率的に状態空間を探索することができる．彼らは，2 台のロボットがいる環境で行動のみ抽象化することで，実時間で協調行動を獲得できることを示した．しかし，複数のロボットがいるような環境では，センサレベルの情報を用い，行動のみ抽象化するだけでは，現実時間では学習が困難である．他者と協調・競合行動を行う際，他者の行動予測が重要となってくる．これは他者の将来の行動が適切に予測可能であれば，他者の行動を考慮に入れた上で，自身のタスク達成に最適な行動決定を行なうことが可能であるためである．Takahashi et al.[6]は，強化学習の枠組を用い，ゴール状態までの距離を表す状態価値を用いて，他者の行動を推定する手法を提案している．この手法では，相手の行動の違いや視点の差による状態認識の違いが存在しても，ある意図に対応する行為の推定した状態価値の増加・減少によって意図を正しく推定できることが確かめられている．センサ

レベルの情報を用いて現在の状況を判断するのではなく，他者行為の状態価値の推定情報を組み込むことで，探索空間が小さくなり，結果的に学習時間が短縮できると考えられる．

本研究では，センサレベルの情報を抽象化した ” 自己行為の状態価値と他者行為の推定した状態価値 ” に基づく協調・競合行動を速やかに獲得する手法を提案する．RoboCup 中型機リーグに出場しているサッカーロボットを想定したシミューレータを用い，5 対 4 でパス，ドリブル，シュートを行うタスクで実験を行ない，本手法の有効性を示す．

2 強化学習

強化学習は，エージェントが環境との相互作用を通して累積報酬を最大化にする枠組である．エージェントは試行錯誤を行いゴールを見つける．状態価値 $V(s)$ とは，ゴール状態までの累積報酬の期待値である (式 1)．ここで簡単な例を示す．Fig.1 では，エージェントが状態 s_0 から s_5 まで移動し，報酬 0 を受け，s_5 で止まる行動をとり，正の報酬 $+1$ を受け取る様子を示している．この経験により，各状態の状態価値 V は s_5 から s_0 へと減衰率 γ の割合で伝播していく．状態価値関数は Fig.2 のようにゴール状態に向かって山のような形の関数となるので，状態価値の高さはゴール状態への近さを反映する．エージェントは状態価値が高くなるような行動をとる．マルチエージェント環境では，センサレベルの情報を使っての学習は探索空間が大きくなり現実時間では困難である．そこで，センサ情報ではなく，センサ情報を抽象化した状態価値を用いることで，探索空間が抑えることを考える．

$$V(s) = E\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\} \qquad (1)$$

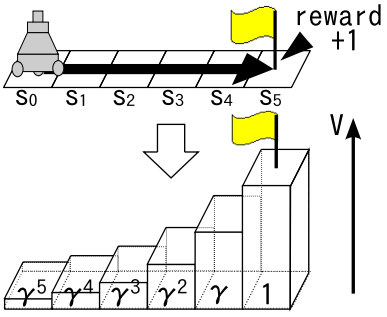

Figure 1: Sketch of state value propagation

3 協調競合行動学習

学習者はいくつかの基本行動モジュールを持っており，これらはあらかじめ強化学習の枠組で状態価値推定と共に獲得されているものとする．また，他者の視点の観測情報を推定し，これを基にある行動の状態価値を推定するモジュールも存在する．この状態価値推定は自己の行為と他者の行為両方に関して同じものを利用する．これらの自己と他者の状態価値推定を基に新たに状態価値空間を張り，協調競合行動学習を行う．

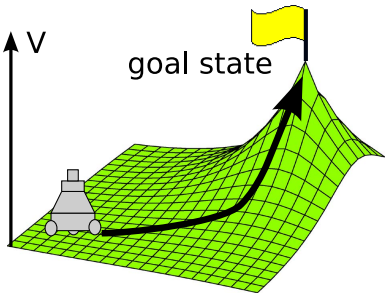

Figure 2: Sketch of a state value function

システムは，下位層に行動モジュール (action module) と他者行為推定モジュール (V estimation module) があり，上位層に学習器 (Gate) があるマルチモジュール型学習機構である (Fig.3)．行動モジュールは観測情報から各行動に対する状態価値を計算する．一方，他者行為推定モジュールは自分の観測情報を基に，自己中心座標系から他者中心座標系への変換を行い，他者の観測情報を推定する．すでに獲得している自分の行動モジュールに他者推定観測情報を当てはめることで，他者の行為の状態価値を推定する．上位層の学習器は，他者行為推定モジュールと行動モジュールから送られてくる状態価値を状態変数として，どの行動モジュールを選択するかを動的計画法の枠組で学習し，選ばれた行動モジュールに従った行動をとる．センサ情報を使って他者の状況を認識するのではなく，他者の行為を推定することによって，探索空間を抑え学習効率を向上させることができる．

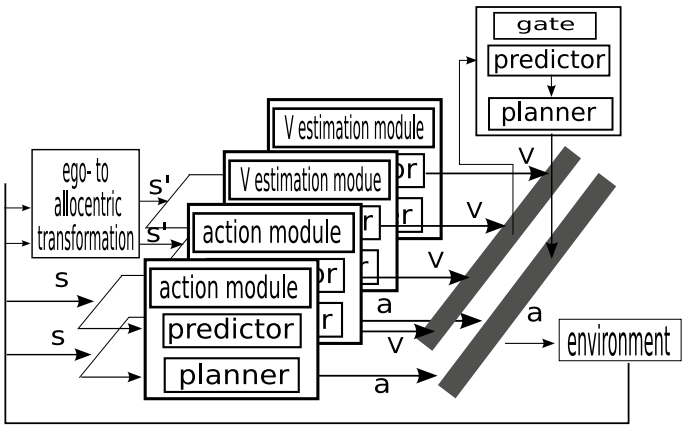

Figure 3: A multi-module learning system

4　タスクと仮定

環境は RoboCup 中型機リーグのフィールドにオフェンスチームが 5 台，デイフェンスチームが 4 台で構成されている．オフェンスチームはパスを回し，シュートをする．ディフェンスチームはオフェンスをマークしながら，ボールが近くにくるとボールをとりにいく．オフェンス (ディフェンス) チームのマーカーの色はマジェンダ (シアン) で，自陣ゴールの色は青色 (黄色) である．ボールに一番近いロボットがパサーとなる．

仮定として，オフェンスチームのパサーのみが学習し，他のレシーバとデイフェンスは固定政策で動くものとする．パサーは基本行動として 4 台のレシーバにパスをするか，ドリブルシュートを用い，状況に応じた適切な行動を学習する．パサーがレシーバに向かって，パスを出した後，パスを受けたレシーバがパサーに，パスを出したパサーがレシーバに切り替わるものとする．１試行が終わるたびに，各ロボットがコミュニケーションを行うことにより，学習情報を共有できるものとする．また，行動モジュールと推定モジュールはあらかじめ，学習しているものとする．

4.1　オフェンスチーム

パサーは他の４台のレシーバのうち，どのレシーバにパスをするか，またドリブル・シュートを選択する．パサーはパスを出した後，ある一定時間だけゴールに向かって移動するパスアンドゴーを行うものとする．レシーバはボールの方を向き，ボールやパサーや他のレシーバに一定距離以上近づかず，長方形を作るように動く (Fig.4)．なお，試行開始時の位置は，自陣でランダムに配置されている．

4.2　デイフェンスチーム

ボールの一番近くにいるオフェンス（パサー）に一番近いデイフェンスがマークをし，残りのディフェンスは一番近いオフェンスをマークする．マークとは，オフェンスの近くでオフェンスと自陣ゴールの間に入ることである (Fig.4)．そして，ボールが近くにくるとボールをとりにくる．また，オフェンスチームの不利にならないように，ペナルティエリアに一定時間入れないものとする．なお，試行開始時の位置は，自陣でセンターサークルに入らないようにランダムに配置されている．

4.3　ロボットと実験環境

RoboCup 中型機リーグに出場しているロボット (Fig.5) を想定したシミュレーションにより実験を行った (Fig.6)．ロボットは，センサに全方位カメラ，前方カメラ，移動機構に全方位移動機構，キック機構を持ったロボットである．Fig.6 の右上は，前方カメラ，右下は全方位カメラの画像を表している．ロボットは，色情報を使って，ボールや他のロボットを認識している．仮定として，ロボットは，全方位カメラから得られるを情報を使って，三次元再構成をし，他者のセンサ情報を推定する．

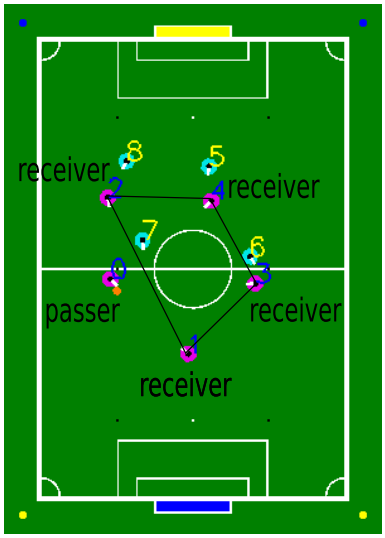

Figure 4: A passer and the defence formation

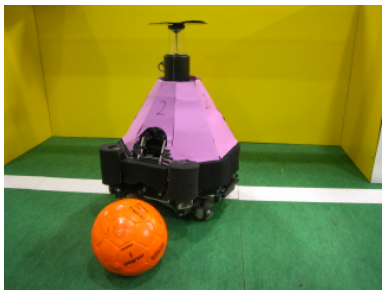

Figure 5: A real robot

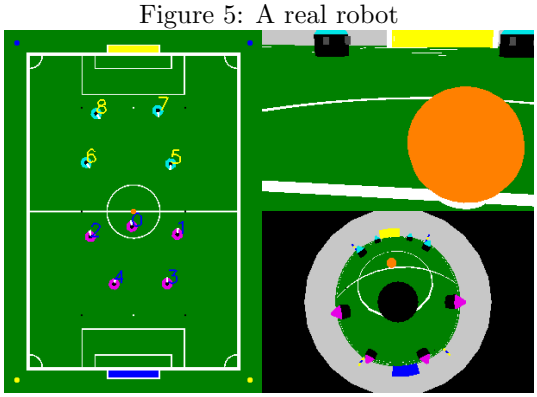

Figure 6: Viewer of simulator

4.4　上位学習器の状態/行動空間

パサー (学習者) の下位層の行動モジュールは，各レシーバに対するパスモジュールを４つ，ドリブル・シュートモジュールを１つ，合計５つある．下位層のレシーバ推定モジュールは，各レシーバーがパスを受けた後のシュートのしやすさの達成度を推定するもので，合計４つある．これらの行動モジュールとレシーバ推定モジュールはあ

らかじめ，上位層の学習の前に獲得しているものとする．下位層の行動空間は，設計者によって設計されたマクロ行動を適用する．マクロ行動は，モータレベルの探索をしないため，探索空間を抑えることができる．パサーの上位層の学習器の状態空間 S は，下位層から送られてくる状態価値から成り立っている．

- 各レシーバに対するパスモジュールの状態価値それぞれ４つ
- ドリブルシュートモジュールの状態価値１つ
- レシーバ推定モジュールの状態価値それぞれ４つ

状態数は，２値化されていて，2^4 x 2 x 2^4=512 である．報酬は次のように与えられている．

- 10 ボールがゴールに入る．(1 試行終了)
- -1 ボールをインターセプトされる．(1 試行終了)
- 0.1 パスが成功する．
- 0.3 ドリブルが成功する．

ボールがフィールドの外に出たり，ある一定時間たつと引きで 1 試行が終了する．以下では，下位層の行動モジュールの詳細を示す．

4.5　パスモジュールの状態空間

パスモジュールの状態空間 S は，全方位カメラ上で，

- レシーバより手前にいるディフェンスの中で，レシーバとのなす角が最も小さいディフェンスとの角度 (θ_1)
- 一番近いディフェンスとレシーバの角度 (θ_2)

である (Fig.7)．二つの角度は，ロボットが見えない状態を含めて，それぞれ 10 個に量子化されている．よって，状態数は 100 である．パスモジュールの状態価値のイメージ図を Fig.8 に示す．Fig.8 はある状況でパサーがパスのしやすさを示したものである．パサーは 3 号機である．各レシーバ (0,1,2,4 号機) の横にあるゲージは各レシーバに対するパスモジュールの状態価値を表していて，ゲージが高いほど状態価値が高い．1，2 号機はディフェンスにパスコースを防がれていないので，状態価値が高い．一方，0，4 号機はディフェンスにパスコースを防がれているので，状態価値が低い．状態価値のマップを Fig. 9 に示す．レシーバとディフェンスの角度が小さい程，状態価値が低い．黒色の領域は未経験の状態である．上位層に送られる状態価値は黄色と赤黒で 2 値化されている．

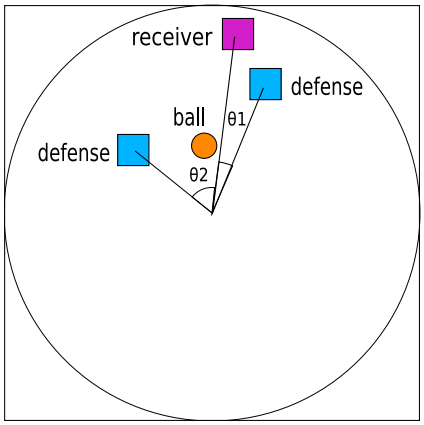

Figure 7: state variable of the pass module

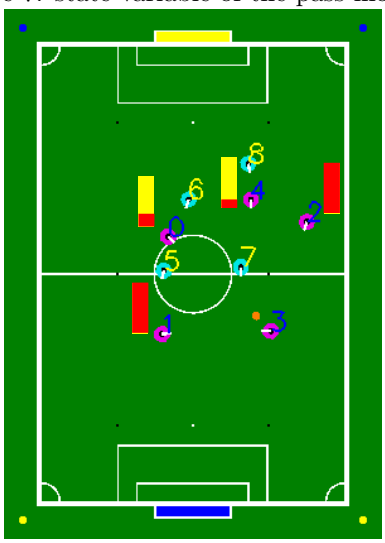

Figure 8: examples of state valuesof the pass module

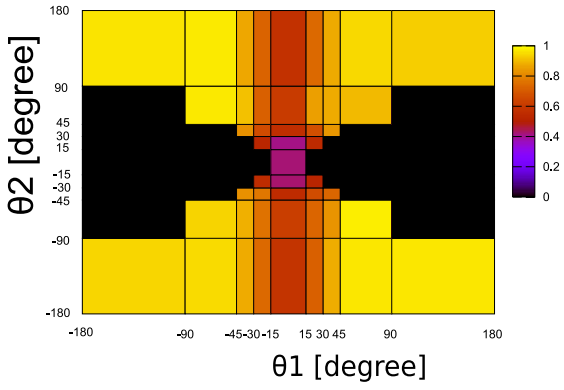

Figure 9: state value map of the pass module

4.6　ドリブル・シュートモジュール

ドリブル・シュートモジュールの状態空間 S は，全方位カメラ上で，

- 一番近いディフェンスと相手ゴールの角度 (θ_1)
- 一番近いディフェンスとボールの角度 (θ_2)
- 一番近いディフェンスの距離 (r)
- 相手ゴールの両エッジの角度 (θ_3)（ゴールまでの距離を表す）

である (Fig.10)．それぞれ，8,8,5,7 に量子化されている．よって，状態数は 8 x 8 x 5 x 7 = 2240 である．パサーのドリブル・シュートモジュールの状態価値のイメージ図を Fig.11 に示す．Fig.11 は，ドリブルシュートのしやすさを示している．左図は，パサー (1 号機) は，ディフェンスが近くにいなく，ゴールに近いので，状態価値が高い．一方，右図は，パサー (3 号機) は，ゴールから遠く，ディフェンスが近くにいるので，状態価値が低い．θ_2 と θ_3 を固定した時の θ_1 と r の状態価値のマップを Fig. 12 に示す．レシーバとディフェンスの角度が小さい程，ゴールから遠い程，状態価値が低い．上位層に送られる状態価値は黄色と赤黒で 2 値化されている．

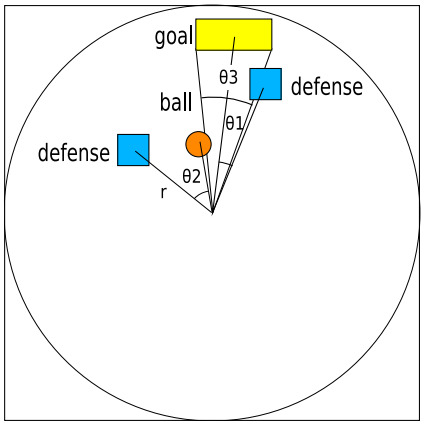

Figure 10: state variables of the dribble and shoot module

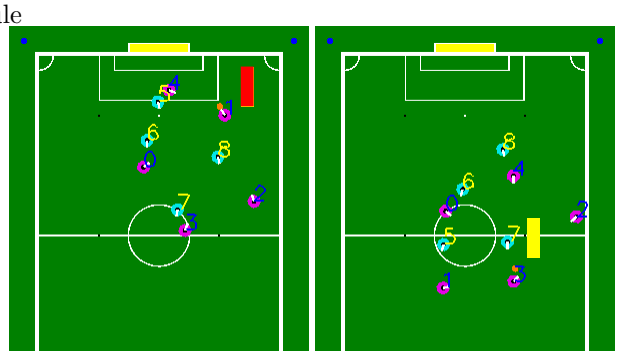

Figure 11: two examples of state values of the dribble and shoot module

4.7　レシーバの推定モジュール

パサーは全方位画像情報から 3 次元再構成をし，レシーバがどのような画像情報を取得しているか計算する．そして，すでに獲得している自分のレシーバモジュールに当てはめてることで，レシーバがパスを受けてからシュートしやすさの推定を行う．レシーバの推定モジュールの状態空間 S は，全方位カメラ上で，

- 一番近いディフェンスの距離 (r)
- 相手ゴールの両エッジの角度 (θ_1)（ゴールまでの距離を表す）

である (Fig.13)．それぞれ，5,7 に量子化されていて，状態数は 5 x 7 = 35 である．レシーバ推定モジュールの状態価値のイメージ図を Fig.14 に示す．Fig.14 に示す．各レシーバ (0,1,3,4 号機) の横にあるゲージは状態価値を表し，状態価値が高いとゲージが高い．0 号機は相手ゴールの近くでディフェンスが近くにいないので状態価値が高い．一方，1 号機は相手ゴールから遠く，ディフェンスが近くにいるので状態価値が低い．レシーバ推定モジュールの状態価値のマップを Fig. 15 に示す．ディフェンスが遠く，ゴールに近いほど，状態価値が高い．黒色の領域は未経験の状態 (ゴールの中に入った) である．上位層に送られる状態価値は黄色と赤黒で 2 値化されている．

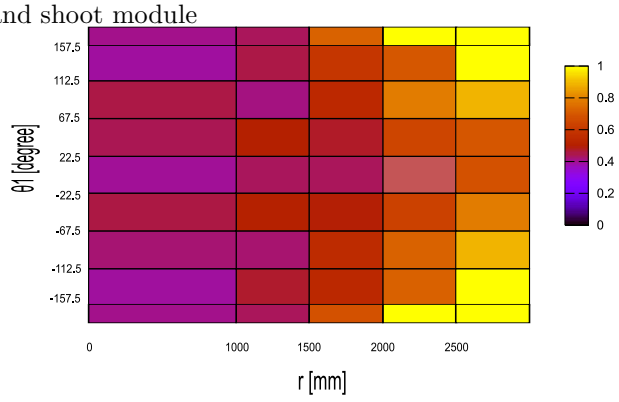

Figure 12: state value map of the dribble and shoot module

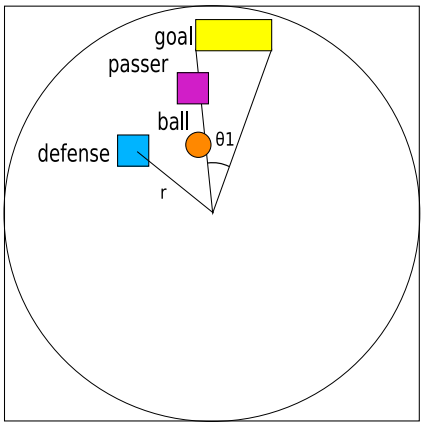

Figure 13: state variables of the receiver module

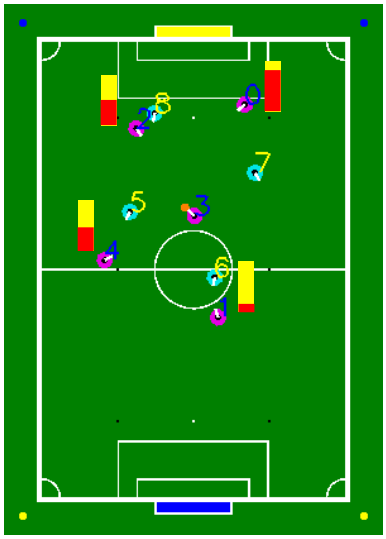

Figure 14: examples of state values of the receiver module

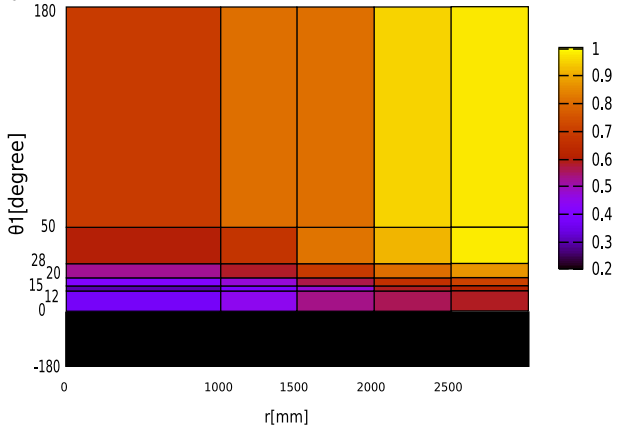

Figure 15: state value map of the receiver module

5 実験結果

成功率を Fig. 5 に示す．下位のモジュール選択を 80% greedy 20%ランダムで行った時のグラフである．900 試行後，成功率が 30%, 失敗率が 70%, 引き分け率が 10%に収束している．Shivaram et al.[3]は，30000 試行後，成功率が 30%程度であり，学習時間は 30 倍程度短くなった．Fig. 16 は，1 試行のパス回数を示している．350 試行以降パス回数が減っている．これは，無駄なパス回しをしていないということである．100%greedy の時の成功率，失敗率，引分率は，それぞれ，55%，35%，10%である．100%random の時は，それぞれ，2%, 97%, 1%である．100%greedy の時の成功率は，80%greedy の時の成功率よりよい．これは，レシーバとディフェンスは固定政策であり，新たな状況がそれほど起こらないからである．獲得された行動の様子の一例を Fig.18 に示す．

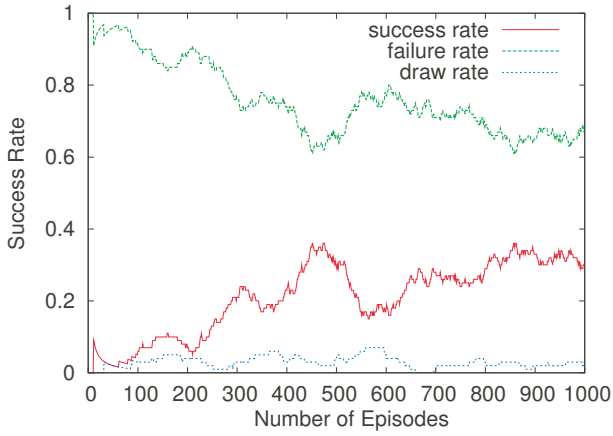

Figure 16: success rate

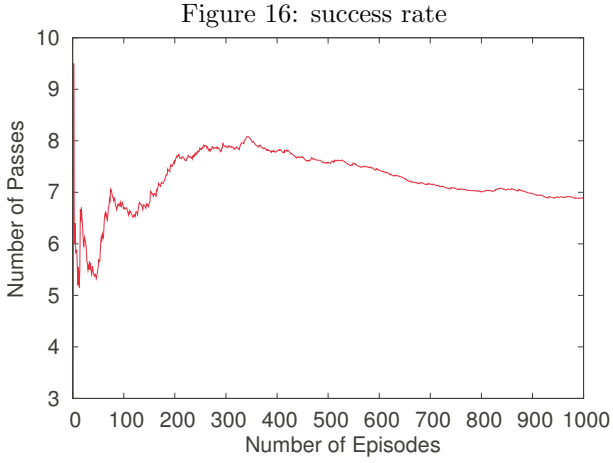

Figure 17: the number of passes

6 考察

学習を加速させるため，センサ情報の代わりに状態価値，モータコマンドの代わりにマクロ行動，そして，レシーバの行動を推定するモジュールを導入した．この結果，学習時間が 30 倍速くなった．比較手法[3]は，30000 試行後，

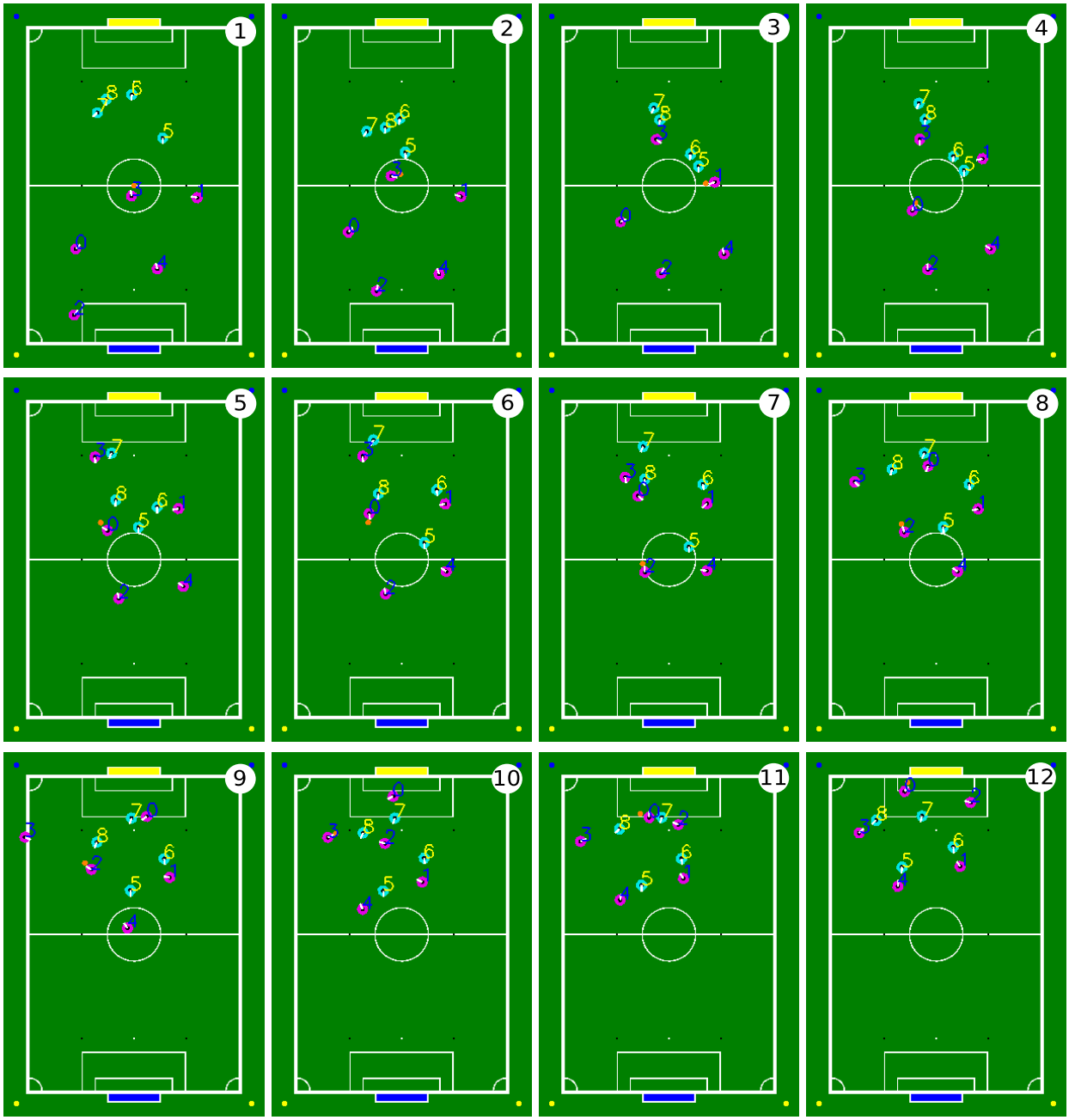

Figure 18: a sequence of a behavior in simulation

成功率がコミュニケーションありで 32%，コミュニケーションなしで 23%に収束している．

本手法では，1 試行中エージェント間でコミュニケーションを行なわないが，レシーバ推定モジュールが同じ役割をしていると考えられる．レシーバ推定モジュールを用いない場合の成功率を Fig. 19 に示す．成功率は 21%程度に収束している．これは比較手法[3]の成功率 23%と近い．状態と行動の抽象化 (状態価値とマクロ行動) は学習時間を抑えることができる．一方で，レシーバ推定モジュールの導入は，チームワークの向上につながる．実機での実験が今後の課題である．

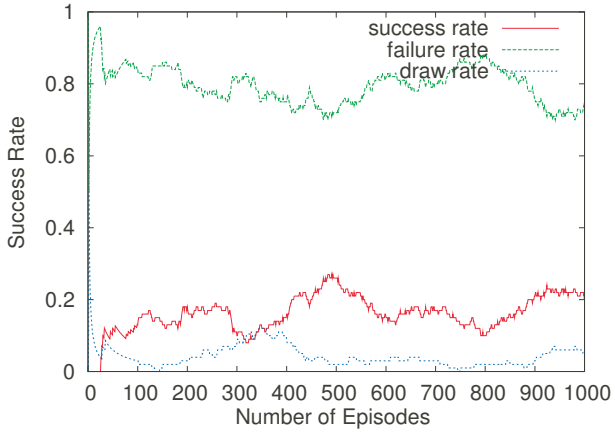

Figure 19: Success rate without the receiver's state inference modules

7 結言

従来，マルチエージェント環境に強化学習を適用する場合，センサレベルの情報を用いて探索すると，状態空間の爆発により現実時間で学習することが困難である問題に直面する．そこで，マルチモジュール学習機構を導入し，センサレベルの情報を抽象化した ” 自己行為の状態価値と他者行為の推定した状態価値 ” を用いて，探索空間を抑えこの問題を解決した．

RoboCup 中型機リーグに出場しているサッカーロボットを想定したシミューレータを用い，5 対 4 でパス，ドリブル，シュートを行うタスクで実験を行ない，本手法の有効性を示した．

参考文献

[1] 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.

[2] 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.

[3] Shivaram Kalyanakrishnan, Yaxin Liu, and Peter Stone. Half field offense in robocup soccer: A multiagent reinforcement learning case study. *In Proceedings CD RoboCup*, 2006.

[4] 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.

[5] Yasutake Takahashi, Kazuhiro Edazawa, and Minoru Asada. Multi-module learning system for behavior acquisition in multi-agent environment. In *Proceedings of 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. CD–ROM 927–931, October 2002.

[6] Yasutake Takahashi, Teruyasu Kawamata, and Minoru Asada. Learning utility for behavior acquisition and intention inference of other agent. In *Proceedings of the 2006 IEEE/RSJ IROS 2006 Workshop on Multi-objective Robotics*, Vol. 1, pp. pp.25–31, 2006.