他者の状態価値の基づく協調・競合行動の獲得

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

○ 野間 健太郎 (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 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.
- 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.

Key Words: reinforcement learning, cooporative/competitive behaviors acquisition, multi-agent system, modular learning system, 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で強化学習を適用し，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)]．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，センサレベルの情報を用いた状況判断を行うと，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.[2)] 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，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．河又 [4)] は，強化学習において，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あっても，ある意図に対応する予測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，いくつかの実験によって確かめられている．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．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，他者の意図情報を状況判断に組み込むことで，探索空間が小さくなり，結果的に学習時間が速くなると考えられる．

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 システム

システムは，下位層には，他者の行為を推定するモジュールと行動モジュールがあり，上位層には学習器があるマルチモジュール型学習機構である (Fig.1)．他者の行為を推定するモジュールは，自分の観測情報を基に，三次元再構成をし，他者の観測情報を推定する．そして，すでに獲得している自分のモジュールに当てはめることで，他者の行為の状態価値を推定する．一方，行動モジュールは，すでに獲得済みで，観測情報から各行動に対する状態価値を計算する．上位層の学習器は，他者の行為を推定する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する．そして，選ばれた行動モジュールに従った行動をとる．

3 タスク

タスクは，5 対 4 で，オフェンスチームはパスを回し，シュートをする．デイ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くるタスクである (Fig.2)．オフェンスチームのマーカーの色はマジェンダで，自陣ゴールの色は青色である．一方，ディ

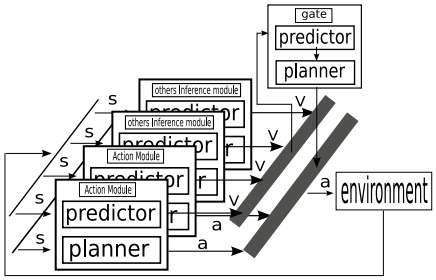


Fig.1 A multi-module learning system

フェンスチームのマーカーの色はシアンで，自陣ゴールの色は黄色である．

仮定として，パサーのみが学習し，レシーバとデイフェンスは固定政策で動いているものとする．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．

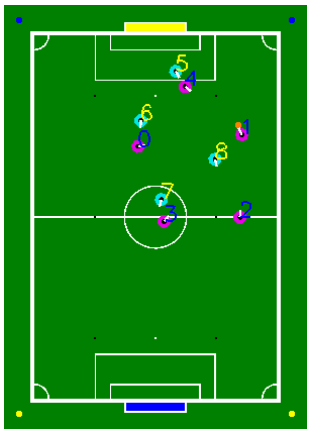


Fig.2 A task: 5 on 4

3.1 オフェンスチーム

パサーは，４台のレシーバにパスをするかドリブル・シュートをする．また，パサーは，レシーバに向かってパスを出した後，ある一定時間だけゴールに向かって移動するようなパスアンドゴーをする．レシーバはボールの方を向いて，両隣の他のレシーバとなす角が９０度になるように動く．また，レシーバは，ボールやパサーや他のレシーバに一定距離以上近づかないように動く．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．

3.2 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）に一番近いがデイフェンスがマークをし，残りのデイ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.2)．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．

3.3 下位層の状態空間

パサーは，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つの行動モジュールを持っている．また，各レシーバーに対する行動の達成度を推定するモジュールを４つ持っている．これらの行動モジュールと他者推定モジュールはあらかじめ，獲得しているものとする．

3.3.1 パスモジュール

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.3(d))．パスモジュールの状態価値のイメージ図を Fig.3(a) に示す．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，状態価値が高いほどゲージが高い．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．

3.3.2 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.3(e))．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.3(b) に示す．パサーは，ディフェンスが近くにいなくシュートをしやすいので，状態価値が高い．

3.3.3 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，レシーバの行動の達成度の推定を行う．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.3(f))．レシーバモジュールの状態価値のイメージ図を Fig.3(c) 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ていて，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．

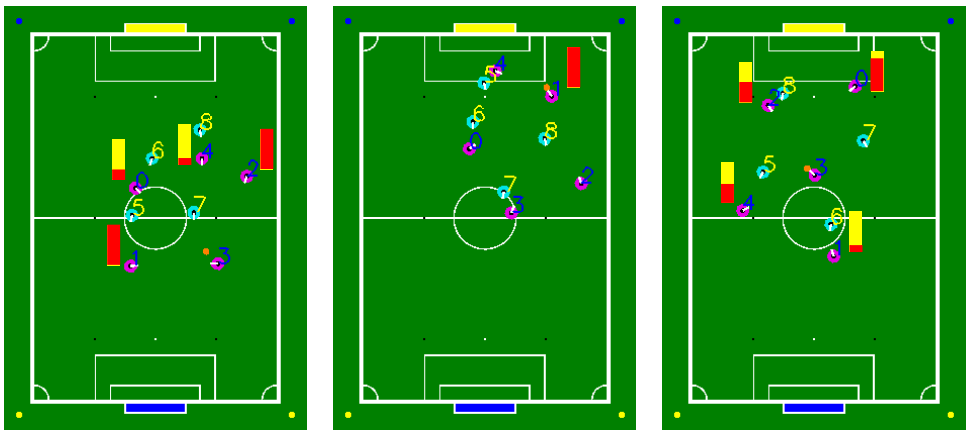
4 実験結果

4.1 獲得された行動の様子

シミュレータにおいて，獲得された行動の様子を Fig.4 に示す．

4.2 タスク成功率とパス回数

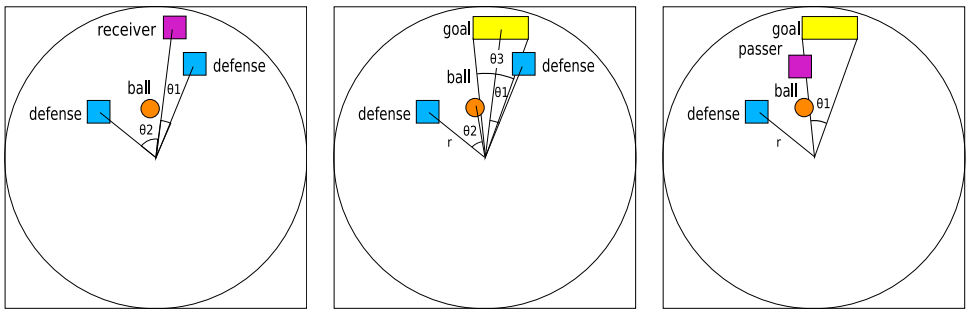
シミュレーションによる学習中のタスク成功率を Fig.5 に，1 試行のパス回数を Fig.6 に示す．また，Table1 と Table2 に 100%greedy と 100%random の結果を示す．



(a) a sketch of state value of pass module

(b) a sketch of state value of dribble/shoot module

(c) a sketch of state value of receiver module



(d) state variables of pass module

(e) state variables of dribble

(f) state variables of receiver

Fig.3 modules of lower layer

Table 1 a success rate in simulation

	成功 [%]	失敗 [%]	引分 [%]
100-greedy	55	35	10
100-random	2	97	1

Table 2 number of pass in simulation

100-greedy	6.5
100-random	4.5

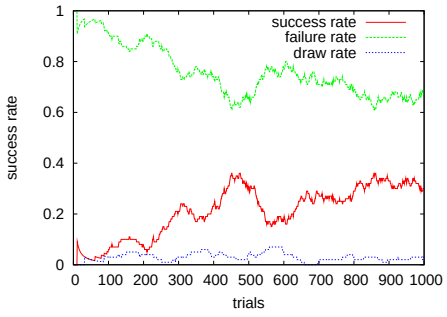


Fig.5 Curves of success rate in simulation

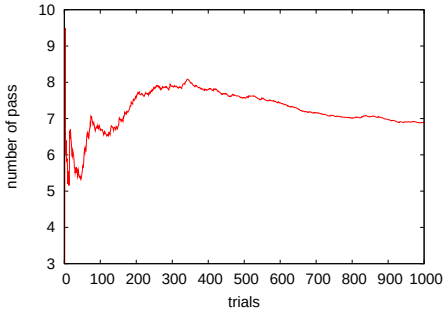


Fig.6 Curves of number of pass in simulation

4.3 タスク成功率と成功率

学習初期は成功率が 10%程度だが，800 試行目で 30%に収束していることがわかる．学習初期は，パスミスが多くパス回数が少ない．350 試行以降，成功率が徐々に上がっているにもかかわらず，パス回数が減っている．これは，無駄なパス回しをせず，できるだけ最短でゴールに到着することができるということである．

4.4 学習時間

図 5 より，約 800 試行で学習が収束していることがわかる．1000 試行の学習時間は，約 1 時間である．センサレベルで学習していた場合の学習時間を計算してみる．パスの状態数が 100，ドリブル・シュートの状態数が 2240，レシーバの状態数が 35 である (3.3 章参考)．よって，状態数は，$100^4 \times 2240 \times 35^4$=$3.36 \times 10^{17}$ である．上位相の状態数 512 で学習時間が 1 時間であるとことを考慮

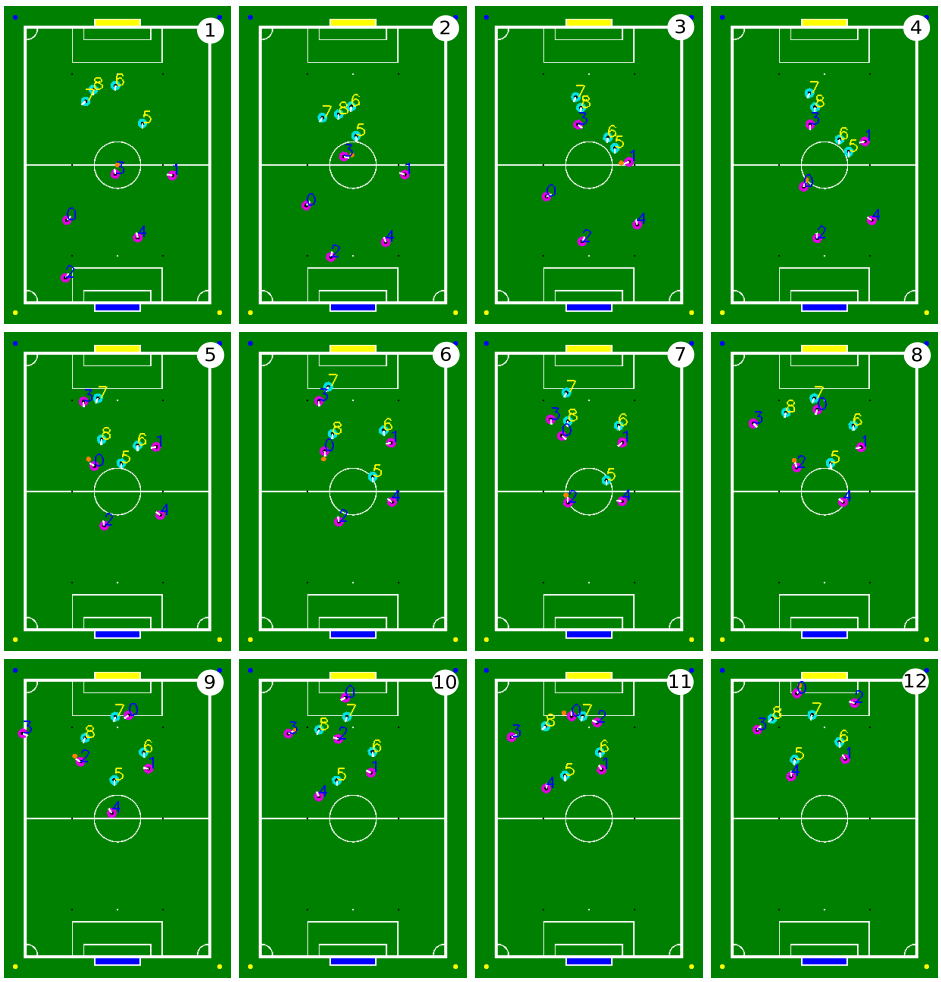


Fig.4 a sequence of a behavior in simulation

すると，センサレベルで学習した場合の学習時間は，6.5 $\times 10^{14}$hour=7.5 $\times 10^{10}$ 年である．現実的な時間では学習できないことがわかる．また，マクロ行動を導入せず，モータレベルの行動を学習した場合を考える．縦，横，回転の組み合わせで行動が決まり，各行動数を 5 とすると，モータレベルの行動を学習する場合の行動数は，5^3=125 である．マクロ行動の行動数は 5 であるので，モータレベルで学習をするとすると探索空間が 25 倍になる．センサレベルの情報を使わずに，状態価値を使って状態を抽象化した場合でも，約 1 日かかることになる．実機で実験を行う場合，試行の際の準備なども含めて，１試行の試行時間はシミュレーションの 100 倍程度かかるとすると，100 日かかることになり現実的に不可能である．

5　結言

本論文では，マルチエージェント環境下で，すみやかに協調・競合行動が獲得できることを示した．

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という問題に直面する．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達成度 (状態価値) に基づく状況認識を行うことで，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] Shivaram Kalyanakrishnan, Yaxin Liu, and Peter Stone. Half field offense in robocup soccer: A multi-agent reinforcement learning case study. *In Proceedings CD RoboCup*, 2006.

[3] 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.

[4] 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.