自己・他者の状態価値推定に基づくチーム行動の生成

Generation of Team Behavior Based on State Value Estimation of Self and Others

○ 島田 皓樹 (阪大院)　　高橋 泰岳 (阪大院)　　正 浅田 稔 (JST ERATO, 阪大院)

Kouki SHIMADA, Graduate School of Eng., Osaka Univ., 2-1 Yamadaoka, Suita, Osaka
Yasutake TAKAHASHI, Graduate School of Eng., Osaka Univ.
Minoru ASADA, JST ERATO, Graduate School of Eng., Osaka Univ.

This paper presents a method that utilizes state value functions of macro actions to explore appropriate behavior efficiently in a multi-agent environment. First, the agent learns a few macro actions and the state value functions based on reinforcement learning beforehand. Second, an appropriate initial controller for learning cooperative behavior is generated based on the state value functions. The initial controller utilizes the state values of the macro actions so that the learner tends to select a good macro action. By combination of the ideas and a two-layer hierarchical system, the proposed method shows better performance during the learning than conventional methods. This paper shows a case study of 4 (defense team) on 5 (offense team) game task, and the learning agent (a passer of the offense team) successfully acquired the teamwork plays (pass and shoot) within shorter learning time.

Key Words: reinforcement learning, state value, multi-agent system, initial controller, RoboCup

1 はじめに

近年，複数のロボットによる協調・競合行動が必要であるサッカーのタスクを取り上げた研究が多くなされている．例えば，Fujii et al. は各ロボットがオフェンス，ボールカバーなど目的行動への達成度を相互通信により共有することにより，各ロボットが役割及び行動を決定する手法を提案した [?]．ロボット自身が各行動に対する達成度を評価することで，置かれている状況を判断し，また，達成度を上げるコントローラを設計することで，目的行動を獲得させている．しかし，行動選択は事前に設計され，固定であるため，環境の変動に対して適応的に行動できない．

そこでロボット自身が試行錯誤を繰り返し，環境や他者との相互作用を通して合目的的な行動を自律的に獲得する強化学習を適用した研究がされている．Kalyanakrishnan et al.[?] はハーフコートのサッカーフィールドで 5 対 4 でパスを行い，シュートを決めるタスクにおいて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，センサレベルの情報によって状況を判断しているため探索空間が大きく，学習時間が長い．そこで，Takahashi et al.[?] はゴール状態までの近さを表す状態価値を状態変数とし，マクロ行動を選択する階層型学習機構を導入することによりセンサレベルの情報とモータレベルの行動を抽象化し，探索空間を抑える手法を提案した．しかし，事前知識なしで学習を始めているため，学習初期段階において探索行動が多い．そこで，状況に応じた協調行動の指針がある方がより効率的に最適行動の獲得ができると考えられる．

そこで本研究では，各エージェントが強化学習により獲得したマクロ行動の状態価値関数を基にチーム全体のゴールへの評価を推定し，これを用いた協調行動生成器と強化学習の組み合わせによりチームの協調行動を効率よく獲得する手法を提案する．マルチエージェント環境下での協調・競合行動は，各エージェントが全体の目的を達成するために行動しており，この複雑な行動は簡単なマクロ行動を組み合わせることで実現できると考えられる．そこで，各マクロ行動の達成度からチーム全体のゴールへの評価値を算出し，ゴールに近づく行動を選択する協調行動生成器を設計する．この協調行動生成器を学習初期段階に利用することで，ランダムな探索行動が減り，チーム行動に有用なマクロ行動を選択しやすくなり，事前知識なしで学習を始めるよりも速やかに協調・競合行動を獲得していくことが期待できる．

2 状態価値推定に基づくチーム行動生成

2.1 実験環境

RoboCup 中型リーグの環境を模したシミュレータ上でのチーム行動学習の獲得実験をする．8m × 12m のフィールドに 5 台のオフェンス，4 台のディフェンスが存在し，オフェンスチームはパスを回して，シュートする．パサーはチームメートにパスをするか，ドリブルシュートし，レシーバはボールの方向を向きながら前後・左右・斜めの 8 方向のいずれかに移動する．一番ボールに近いオフェンスがパサーとなり，パサーがレシーバに対してパスをすると，ボールを受け取ったレシーバがパサーに，ボールを渡したパサーがレシーバへと役割が切り替わっていく．ディフェンスはオフェンスをマークしながら，ボールが近づくとボールを取りにいく．本実験ではパサーのみが行動選択を学習し，ディフェンスチームは事前に設計された方策に基づき行動する．

2.2 システム概要

各エージェントは 2 階層からなるモジュール型学習機構を持つ (Fig.1)．下位層には観察から他者の行為の達成度を推定するモジュールと自己の行為モジュールがある．上位層では下位層の各モジュールから送られてくる情報に基づき状況判断をし，行為モジュールの選択をする協調行動生成器と学習器がある．下位層の自己・他者の行動達成度の情報を状態変数として，どの行為モジュールを選択する

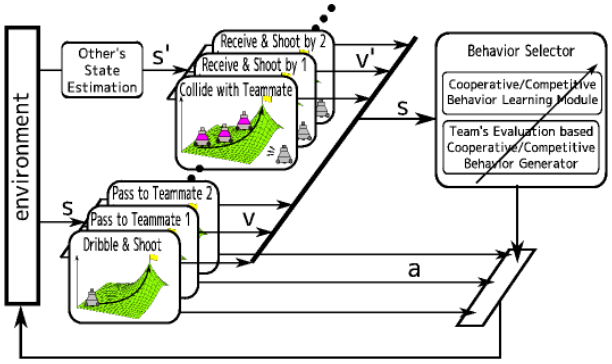


Fig.1 A hierarchical modular learning system

かを学習する．未経験の状態にいる場合に，下位層の情報を基にした協調行動生成器を初期コントローラとして用いて行動する．

2.3 下位層の設定

各エージェントは事前にドリブルシュート，パスなどの基本行動を関数近似手法として CMAC を用いた Q 学習により獲得する．獲得済みの行為に関する状態価値関数を利用することにより，自己・他者の状態価値を推定し，行為の達成度を推定する．各エージェントが持つ下位層のモジュールは以下のとおりである．

2.3.1 ドリブルシュートモジュールに関するモジュール

パサーのドリブルシュートモジュールの状態変数を全方位カメラ上で，

- ボールとの距離
- ゴールとの距離
- ボールとゴールのなす角度

を 3 次元再構成した値と設定する (Fig.2(a))．ボールがゴールに入った時のみ報酬 +1 が与えられ，それ以外は報酬は与えられない．

強化学習により獲得された状態価値関数を Fig.2(b) に示す．この状態価値関数は，ゴールとの距離とボールとゴールのなす角度の関係を表し，ボールとゴールのなす角度がゼロに近づく，つまりそれらが一直線上に並ぶほど，また，ロボットがゴールに近づくほど状態価値が高くなる．

また，ディフェンスプレイヤーがゴールの前にいる場合はシュートが決まらないので，ドリブル・シュートの成功を推定するモジュールを設計する．状態変数を全方位カメラ上で

- ゴールの手前にいるディフェンスプレイヤーの中で，一番ゴールとのなす角が小さいディフェンスプレイヤーとの角度

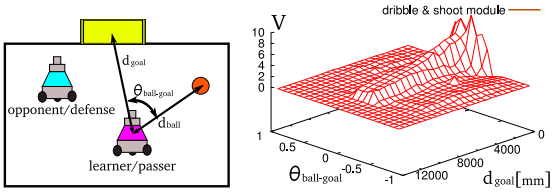
と設定する (Fig.3(a))．ディフェンスにボールが触れたら報酬 -1 が与えられる．

強化学習により獲得された状態価値関数を Fig.3(b) に示す．この状態価値関数より，ディフェンスプレイヤーとゴールのなす角度がゼロに近づく，つまりそれらが一直線上に並ぶほど，状態価値が低くなることがわかる．

2.3.2 パスに関するモジュール

パサーのパスモジュールの状態変数を全方位カメラ上で，

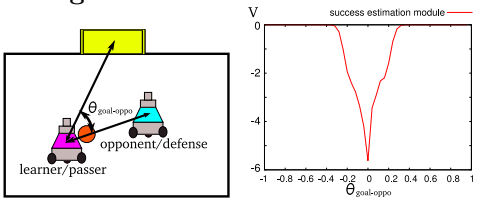
- レシーバの手前にいるディフェンスプレイヤーの中で，一番レシーバとのなす角が小さいディフェンスプレイヤーとの角度



(a) State variables for dribble & shoot module

(b) State value function of dribble & shoot module

Fig.2 A dribble and shoot module



(a) State variables for success estimation module

(b) State value function of success estimation module

Fig.3 A success estimation module

と設定する (Fig.4(a))．ディフェンスプレイヤーにボールが触れたら報酬 -1 が与えられる．

強化学習により獲得された状態価値関数を Fig.4(b) に示す．この状態価値関数はドリブルシュートの成功推定モジュールの状態価値関数 Fig.3(b) と同じ形状である．これは，ボールを蹴る対象が異なるだけで，ディフェンスと対象の角度という状態変数の設定，ディフェンスとボールが触れたら与えられる報酬の設定が同じためである．

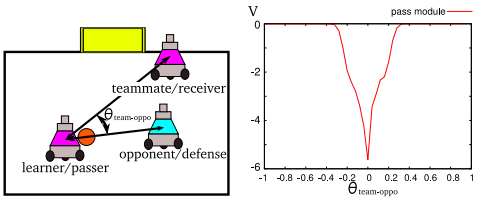
2.3.3 チームメートとの衝突に関するモジュール

レシーバの衝突モジュールの状態変数を

- 一番近いチームメートとの距離

を 3 次元再構成した値と設定する (Fig.5(a))．レシーバが他のチームメートと衝突すると報酬 -1 が与えられる．

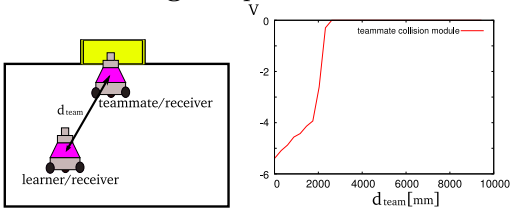
強化学習により獲得された状態価値関数を Fig.5(b) に示す．この状態価値関数より，チームメートとの距離が近づくほど状態価値が低くなることがわかる．



(a) State variables for pass module

(b) State value function of pass module

Fig.4 A pass module



(a) State variables for teammate collision module

(b) State value function of teammate collision module

Fig.5 A teammate collision module

2.3.4 レシーブに関するモジュール

パサーは自身の獲得した行為モジュールのドリブルシュートモジュールを用いて各レシーバがボールを受け取った場合，どれだけシュートしやすいかを推定する．

まず，パサーは各レシーバがボールを受け取った状態を推定するために，レシーバとゴールの間に仮想的にボールがあると仮定する．このとき，レシーブモジュールの状態変数は，

- レシーバと仮想的なボールの距離
- レシーバとゴールとの距離
- 仮想的なボールとゴールのなす角度
- レシーバとゴールの間にいるディフェンスの中で，一番ゴールとのなす角が小さいディフェンスとの角度

を 3 次元再構成した値と設定する (Fig.6)．これらの状態変数を用いて行為モジュールの一つであるドリブルシュートモジュールの状態価値関数 Fig.2(b)，Fig.3(b) を使用して状態価値を推定する．これにより，レシーバがボールを受け取ったときのシュートしやすさの評価値を獲得でき，レシーブモジュールの評価値として設定する．今回の実験では，$d_{ball(virtual)} = 600$ [mm]，$\theta_{ball(virtual)-goal} = 0.0$ としている．

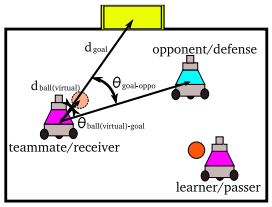


Fig.6 State variables for receive module

2.4 上位層の協調行動学習器

センサレベルの状態変数，コマンドレベルの行動により状態空間，行動空間を作成すると探索空間が大きくなる．そこで上位層の学習器では，基本行動に関する行為モジュール，推定モジュールからなる下位層のモジュールの状態価値を状態変数として利用し，どの行為モジュールを選択するかを強化学習の枠組みで学習する [?]．

パサーの行動選択の学習では，状態変数を

- 推定される各レシーバのレシーブモジュールの状態価値，4 つ
- ドリブルシュートモジュールの状態価値，1 つ
- 各レシーバへのパスモジュールの状態価値，4 つ

と設定する．状態空間は 9 次元で，状態数はそれぞれ，$2^4 \times 2 \times 2^4$ で合計 512 である．8 枚のタイルを設定した CMAC を用いた Q 学習により学習する．

マクロ行動としてパサーは，各レシーバへのパス及びドリブルシュートを事前に獲得済みであり，この 5 通りの行動から選択をする．パサーに与えられる報酬は，ボールがゴールに入った場合，+1，ボールをディフェンスチームにとられた場合，-1 と設定する．

2.5 上位層の協調行動生成器

複数エージェントによる協調行動のような複雑な行動はタスク分解ができ，複数のサブゴールを達成していくマクロ行動の組み合わせだと考えられる．そこで，マクロ行動に関する下位層のモジュールから送られてくる評価値に基づいたチーム全体のゴールへの達成度を算出する．

学習者であるパサーが推定する各ロボットのチーム全体のゴールへの評価値 E を式 (1) に示す．オフェンスチームにおけるチーム全体の目標はボールをゴールの中に入れて得点することである．まず，パサーは自己の下位層のモジュールの状態価値から自身のチーム全体のゴールへの評価値を計算する．下位層のモジュールとしてドリブルシュートモジュールの状態価値 $V_{Dribble\&Shoot}$ と，蹴ったボールが敵に取られずにゴールに入るかどうかを推定するモジュールの状態価値 $V_{SuccessEstim}$ を用いることで自己評価値を算出する．

次にパサーは各レシーバのゴールへの評価値を推定する．推定されるレシーブモジュールの状態価値 $V'_{Receive(i)}$ とパサーのパスしやすさを表すパスモジュールの状態価値 $V_{PassBall(i)}$ の線形和からチーム全体のゴールへの評価値を推定する．

パサーの自己評価による評価値と推定される他レシーバの評価値を比較し，評価値が一番大きいロボットが次にボールを保持するようにパサーはマクロ行動 ma を選択する (式 (2))．パサーの評価が一番高ければボールを保持したままドリブルシュートをし，レシーバの評価が一番高ければそのレシーバに向かってパスする．これにより，チーム全体のゴールへの評価値が上がる行動を選択することができ，ゴール状態に近づくことができる．今回の実験ではパサーにおけるチーム全体のゴールへの評価値に用いる重みとして $\beta = 0.5$ を使用した．

$$E_i = \begin{cases} V_{Dribble\&Shoot} + \beta V_{SuccessEstim} \\ \quad (\text{if } i = \text{self id}) \\ V'_{Receive(i)} + \beta V_{PassBall(i)} \\ \quad (\text{if } i \neq \text{self id}) \end{cases} \tag{1}$$

$$ma = \arg\max_i E_i \tag{2}$$

3 実験結果と考察

3.1 実ロボットによる 2 台対 1 台での協調行動生成器を用いた実験

チームのゴールへの評価値に基づく協調行動生成器の有効性を示すため，実ロボットに実装し，2 台対 1 台の環境でパサーの行動選択実験をした．本実験では，下位層のマクロ行動の状態価値はシミュレーションにより獲得されたものを使用している．

ディフェンスの位置がパサーの正面の場合と，ディフェンスの位置がレシーバの正面の場合での各ロボットの行動の様子を Fig.7, Fig.8 に示す．パサーの前にディフェンスがいる状況では，レシーバの方がチーム全体のゴールに近いとパサーが判断し，パスをしている (Fig.7)．また，レシーバの前にディフェンスがいる状況ではパサー自身の方がチーム全体のゴールに近いと判断し，ドリブルシュートをしている (Fig.8)．

これより，パサーがパサー自身とレシーバのどちらがゴールへ近いか評価し，ドリブルシュート行動とパス行動を正しく選択していることがわかる．

3.2 5 台対 4 台でのチーム行動獲得実験

3.2.1 タスク成功率

ボールがゴールに入った場合を成功と設定し，タスク成功率を求める．100 試行毎の移動平均によって求めた協調行動生成器を初期コントローラとして導入したときの学習中のタスク成功率と初期コントローラを導入せずに学習をしたときのタスク成功率を Fig.9 に示す．初期コントローラを用いることにより学習時間の短縮ができているといえる．

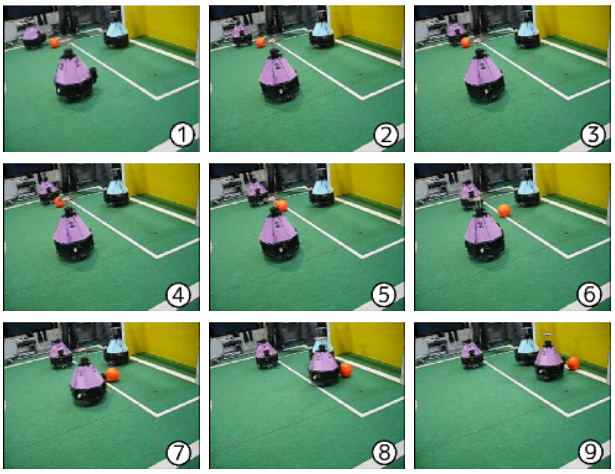


Fig.7 A sequence of behavior when defense is in front of passer

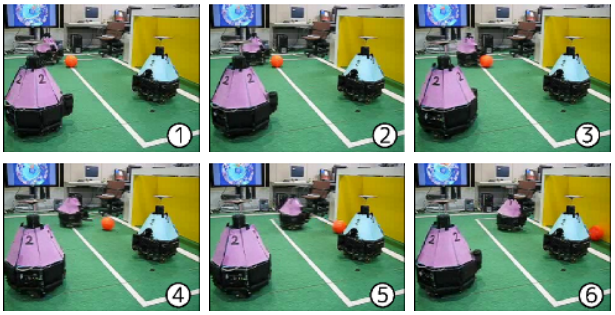


Fig.8 A sequence of behavior when defense is in front of receiver

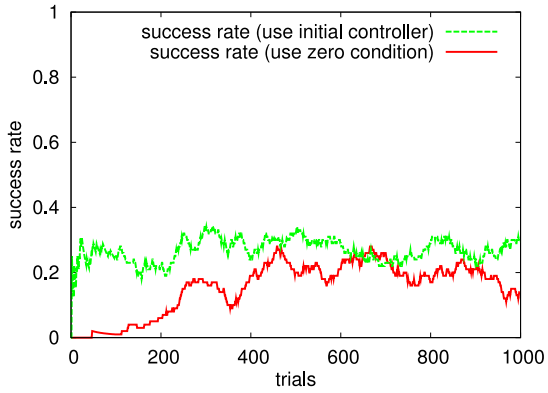


Fig.9 Success rates

3.2.2 獲得されたチーム行動

学習を通じて獲得されたチーム行動の様子を Fig.10 に示す．パサー，レシーバが自律的に動きながら，パス，ドリブルをすることでゴールにシュートしていることがわかる．レシーバの動きは各自がパスの受けやすい方向，シュートの決めやすい方向を判断しながら動いていき，結果的に全体として敵陣に向かっていることがわかる．

4 まとめ

本論文では，マルチエージェントの協調・競合行動を複数のサブタスクに分解し，サブタスクを個々に学習させた後に協調・競合行動を学習する階層型学習機構を導入し，

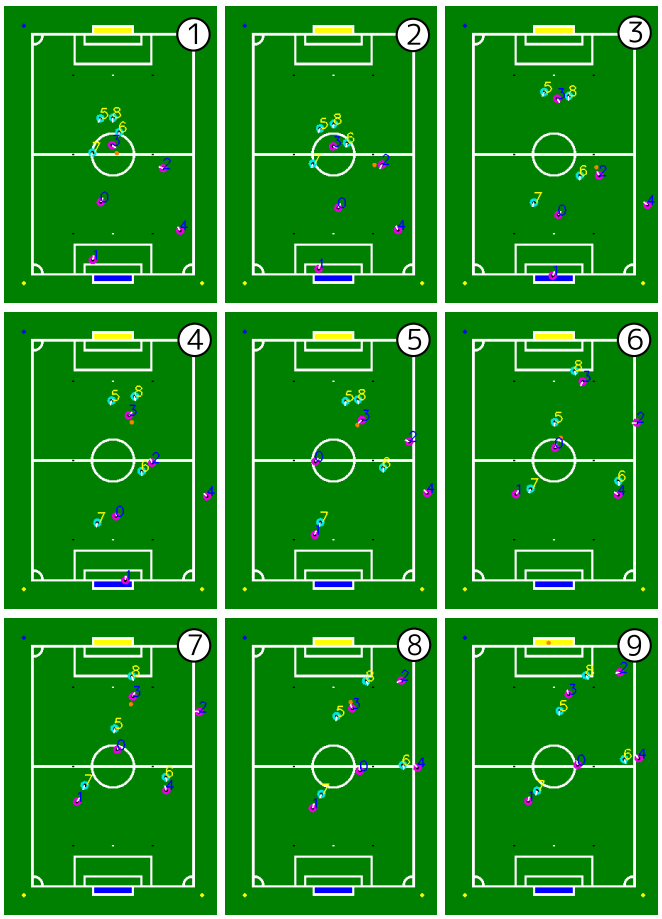


Fig.10 A sequence of acquired behavior

チーム全体の評価に基づく協調行動生成器を設計し，学習初期段階に適用する手法を提案した．本手法では，学習初期にランダムな探索行動を試す必要や，設計者による細かい報酬の設計の必要がなく，効率的なチーム行動の学習が可能である．

この手法を 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レータを用いてフィールド内で 5 対 4 でパスを行い，シュートを決めるタスクで実験を行い，従来手法と比べて学習時間が短時間で協調行動が獲得されることを示した．

参考文献

[1] Hikari Fujii, Masayuki Kato, and Kazuo Yoshida. Cooperative action control based on evaluating objective achievements. In Ansgar Bredenfeld, Adam Jacoff, Itsuki Noda, and Yasutake Takahashi, editors, *RoboCup2005: Robot Soccer World Cup IX*, pp. 208–218, 2005.

[2] Shivaram Kalyanakrishnan, Peter Stone, and Yaxin Liu. Model-based reinforcement learning in a complex domain. In *Proceedings of the 11th annual RoboCup International Symposium*, 2007. CD-ROM.

[3] Yasutake Takahashi, Kentaro Noma, and Minoru Asada. Efficient behavior learning based on state value estimation of self and others. In *Advanced Robotics*, Vol. 22, pp. 1379–1395, 2008.