A4-2 ファジィ状態分割型 Shaping 強化学習を用いた自律移動ロボットの行動獲得に関する研究

福井大学大学院 工学研究科 知能システム工学専攻 進化ロボット研究室
長谷川 大樹 (指導教員：前田 陽一郎)

1. 緒 言

近年、強化学習に代表される外部環境との相互作用を通して試行錯誤によりロボットのコントローラを構築するアプローチの有効性が認知されている [1]。強化学習は学習過程と実行過程を明示的に区別しないオンライン学習であり、未知環境下におけるロボットの行動学習等に広く適用されている [2]。しかしながら、他の学習手法と比較すると強化学習は非常に多くの学習時間を必要とするといった問題も知られている。そこで、これらの問題を解決するために近藤ら [3] は未知環境下での探索において逐次的に得られる入出力データと外部評価から、NGNet を用いてロボットのコントローラを関数近似し、入出力の静的写像を構築するための Actor-Critic 法による結合荷重パラメータ学習手法を提案している。

また、一般にロボットに効率よく学習させるために生物の学習メカニズムから工学的模倣を行うのは有効な方法である。田渕ら [4] は他者の行動を真似ることで大まかに望ましい行動に近づき、次いで副報酬を伴った学習を行うことで模倣で得た方策を洗練していく手法を提案している。さらに、動物の調教やトレーニングなどに用いられる Shaping の概念をロボットの行動学習に応用する研究も行われており、その有効性が検証されている。

本研究室ではこれまで、階層型ファジィ行動制御、およびこの上位の行動選択ファジィルールを強化学習により自動獲得する手法 [5] を提案してきた。さらに学習時の膨大な試行回数を削減するために Shaping 強化学習 [6] を用いてこれまで考慮していなかったロボットと人間との関わりなどの外界からのインタラクションの設計をすることによるロボットの効果的な学習手法を提案してきた。しかしながら、これらの強化学習を用いた研究はグリッド探索問題のような比較的単純な行動学習でしか検証されていなかった。また Shaping 学習は調教者が逐一ロボットを観察して報酬を与えなければならないためユーザに多大な負担がかかってしまうという問題がある。

そこで本研究では、より複雑なサッカーロボットの行動獲得を例題とし、Shaping における調教者の報酬授与特性を基に自動化を行ったファジィ状態分割型 Shaping Profit Sharing 法（SFPS）を提案する。本論文では階層型ファジィ行動制御の上位の行動選択ファジィルールを学習する強化学習に Shaping の概念を取り入れることによってロボットに段階的な目標行動学習を行わせる。さらに Shaping 学習を行うにあたってユーザにかかる負担を軽減するために、報酬授与特性を基に自動化したシステムを組み込む。本手法の有効性を検証するため、RoboCup 中型ロボットリーグ規格に基づいたシミュレーション実験により学習性能評価を行ったので、その結果について報告する。

2. Shaping について

Shaping は犬やイルカなどの動物の調教に用いられる用語であり、行動分析学の分野でも効率的に行動を強化するための有効な概念として注目されている。Shaping とは行動を形成するという意味であるが、少しずつ行動を強化しながら目標行動に近づけていく概念である。

3. ファジィ状態分割型強化学習を用いた行動獲得手法

本研究ではファジィ状態分割型強化学習 (FPS) を基に、Shaping の概念を導入した手法 (SFPS)、Profit Sharing の学習パラメータをファジィルールによってチューニングした手法 (FASSFPS)、Shaping 報酬付与を自動化し分化強化した手法 (DR-FASSFPS) を提案する。

3·1 ファジィ状態分割型 Shaping 強化学習 (SFPS)

図 1 に本手法で提案した FPS を用いた階層型ファジィ行動制御システムの概念図を示す。本手法ではまず、学習に用いる入力情報としてセンサなどから得られる環境情報をそのまま用いるのではなく、その前段階でマクロ状態認識ファジィルールを用いて環境状態をマクロ化することによって状態の次元数を減らしている [5]。

学習開始時には、すべての後件部シングルトンの初期値はある一定値を与えているのでどのサブタスクも同一の確率で選択され、ルーレット選択を行い探索する。そして、報酬 r_t が得られた場合、報酬が得られた時点を 0 ステップと考え、式 (1),(2) より h ステップ前の行動 i を取った場合の行動重みを更新する。

$$f(h) = \gamma^h (r_t + S(t)) \tag{1}$$

$$w_{ik} \leftarrow w_{ik} + \alpha f(h) \tag{2}$$

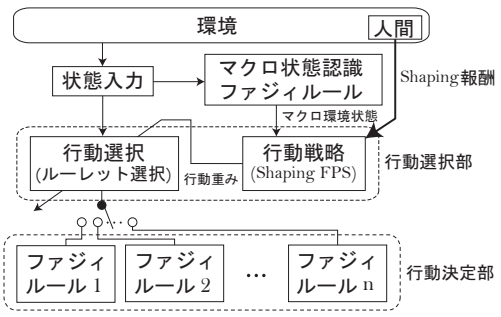

図 1: SFPS を用いた階層型ファジィ行動制御システムの概念図

ω_{ik} : 行動重み　$f(h)$: 強化関数
α : 学習率　γ : 割引率　r_t : 報酬
$S(t)$: Shaping 報酬　$S(t) = c$ (c は定数)

さらにここで調教者が適宜 Shaping 報酬 $S(t)$ を与え、人為的に報酬を変化させることにより、報酬 r_t が得られた行動に加え Shaping 報酬が与えられた行動も重要と認識し、別途報酬を追加で与えることにより、連続した行動系列を効率よく学習させる。尚、以下のファジィルールは紙面の都合により割愛する。

3·2　行動決定ファジィルール

行動決定ファジィルールは、最終的にロボットが行う個別タスクを表現したファジィルールであり、表 1 に示すようにボール追跡やドリブルなどの独立したロボットの行動を簡略化ファジィ推論で記述している。

表 1: 行動決定ファジィルールの入出力変数

観測される環境情報 (入力)	ロボットの基本行動 (出力)
ボールの相対距離、相対方位	ボール追跡
ゴールの相対距離、相対方位	ドリブル
ゴールの相対距離、相対方位	シュート
障害物の相対距離、相対方位	障害物回避

3·3　マクロ状態認識ファジィルール

表 2 に示すような観測された環境情報を一旦マクロ環境状態に落とし、これと獲得した報酬や、Shaping 報酬を入力として Shaping FPS により学習する。学習後には、あるマクロ環境状態における行動選択ファジィルールの行動重みが獲得される。

表 2: マクロ状態認識ファジィルールの入出力変数

観測される環境情報 (入力)	マクロ環境状態 (出力)
ボールの相対距離、相対方位	ボール追跡度 (Dc)
障害物の相対距離、相対方位	障害物回避度 (De)
ゴールの相対距離、相対方位	シュート決定度 (Dd)

3·4　行動選択ファジィルール

マクロ状態認識ファジィルールにより、上位の強化学習に用いる状態次元数を減少させることができたので、これらを用いて行動の重み付けを行う (表 3)。本手法ではファジィルールで出力した行動重みを基に、ルーレット選択により下位行動の選択を行う。この行動重みは前述の式 (1),(2) により更新され、学習が進むにつれて最適な行動を獲得するようになる。

表 3: 行動選択ファジィルールの入出力変数

マクロ環境情報 (入力)	各行動重み (出力)
ボール追跡度 (Dc)	ボール追跡の行動重み
障害物回避度 (De)	障害物回避の行動重み
シュート決定度 (Dd)	シュートの行動重み

3·5　ファジィ適応型探索 FPS の導入

未知の環境に対する学習では、学習初期の探索領域が大きい場合では積極的な探索を行い、学習が進み探索領域が絞られた段階では既知情報を生かした学習が求められる。そこで本研究では FPS 中の Profit Sharing の学習パラメータをファジィルールにより適宜調整を行い、状況に応じて学習を行うファジィ適応型探索 FPS(FASFPS) を追加提案する。

図 2 に本手法のアルゴリズムフローを示す。Profit Sharing は等比減少関数を用いて行動重みの更新を行っていくが、このとき用いる学習率 α、割引率 γ をファジィルールによってチューニングを行う。

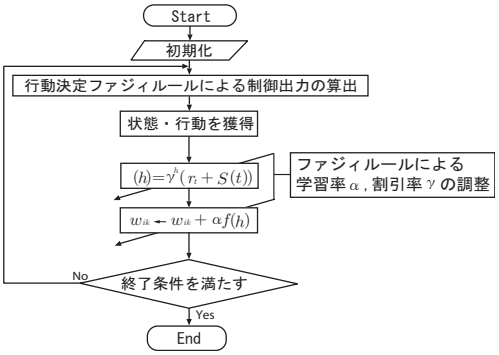

図 2: FASFPS のアルゴリズムフロー

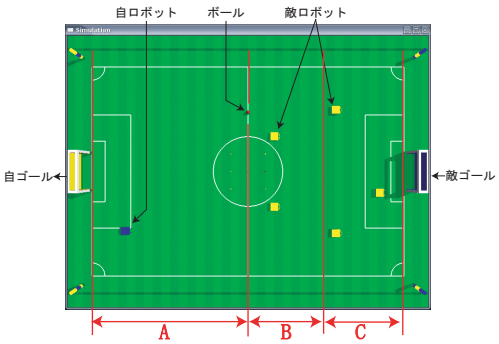

図 3: 実験で用いたロボットの初期配置

3·6　Shaping の自動化

Shaping は調教者がロボットを観測して適宜与えるものであるため、調教者に大きな負担がかかるといった問題がある。また、これまでの学習では目標行動までの中で学習の進行過程に関係なく同じように Shaping 報酬を与えていたが、これでは問題が複雑になるほど Shaping 報酬を与える回数が多くなり効率が悪くなる。そこで、調教者の行った報酬分配傾向から Shaping 報酬付与を自動化し、3·5 節で述べた FASFPS で得られた知見を基に、より複雑な環境でも効率よく学習することのできる分化強化型 FAS-ShapingFPS(DR-FASSFPS) を提案する。DR-FASSFPS の学習手順を以下に示す。

Step1 あらかじめ図 3 に示すようにサッカーフィールドを区切り、学習を行う自ロボットから近い順にエリアを A,B,C とする。

Step2 最初の学習は前述した Shaping FPS を用いて行い、学習で調教者 (人間) がロボットに与えた報酬の記録を行う。Step1 で区切った A,B,C の範囲別に記録を行うため合計で 9 個のファイル（27 個のグラフ）が作成される。

Step3 Step2 で取得したデータを基にグラフを作成し、範囲・行動別での Shaping 報酬の与えられた状態の抽出を行う。

Step4 Step3 で得られた抽出結果を基に自動学習を行う。

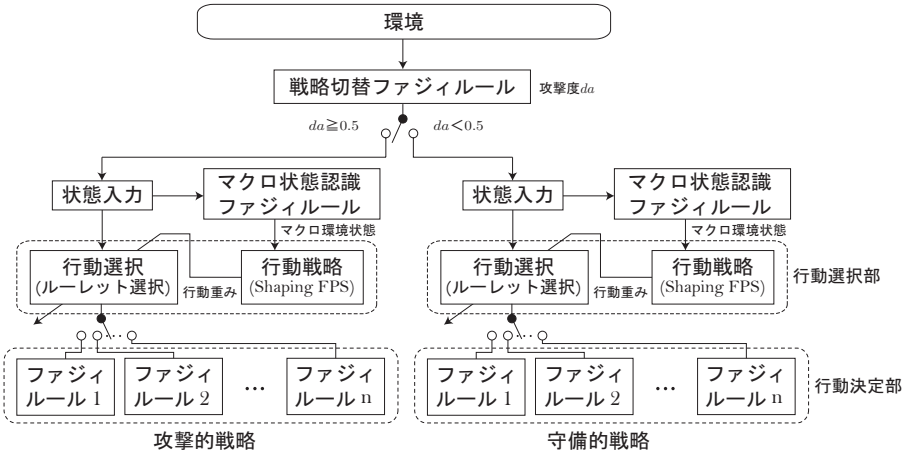

図 4: Shaping FPS による戦略切替手法の概念図

4. SFPS による戦略切替手法

前章までの学習ではサッカー行動の攻撃部分に関する行動決定、行動選択といった、いわゆる戦術レベルの学習までしか行われていなかった。そこで本章では攻撃的戦略と守備的戦略とを上位のファジィルールによって重み付けした戦略切替手法を提案する。

図 4 に本研究で提案する SFPS による戦略獲得手法の概念図を示す。ここでは 3 章で提案した SFPS を攻撃的戦略部、守備的戦略部に分け、上位の戦略切替ファジィルールによって重み付け学習を行う。本手法ではまず、戦略切替ファジィルールによって攻撃度 da を算出する。戦略切替ファジィルールは入力として自ロボットとボールの距離、敵ロボットとボールの距離情報を用いた。これによって今の状態が攻撃行動に向いているのか守備行動に向いているのかを認知させ、攻撃行動を促す度合い (攻撃度 da) を推論する。攻撃度が大きければ (本研究では $da \geq 0.5$ とした) 攻撃行動を学習し、小さければ ($da < 0.5$) 守備行動を学習するように設定した。

この戦略切替手法でも前章と同様に調教者が適宜 Shaping 報酬を与えることによって、その行動を重要と認識し効率よく学習させる。

5. シミュレーション実験

3 章、4 章で提案した本手法の有効性を検証するためにサッカーロボットシミュレータを開発し、シミュレーション実験を行った。シミュレーション画面を図 5 に示す。

ロボットはサッカーの行動を各強化学習法を用いて学習するが、ロボットに Shaping 報酬を与える場合、調教者がフィールド上のロボットの動きを観察して、Joypad の各ボタンを使って Shaping 報酬を適宜与える。以下に本実験の条件を示す。

- フィールドの大きさは 12m × 18m とする。
- ロボットは全方位移動機構、全方位カメラ、キック機構、ボール保持機構を搭載している。
- 自ロボット、敵ロボットの初期位置および台数は実験によって異なる。

実験でロボットに与える報酬はロボットが以下に示す条件を満たした場合に与えた。

- ゴール報酬：ゴールにシュートして得点を入れた時
- 追跡報酬：ボールに近づいた (1m 以内) 時
- 回避報酬：危険地帯 (障害物までの距離が 1.5m 以内) から脱出した時
- 罰：障害物 (敵ロボットやフィールドを囲う壁) に衝突した時
- ブロック報酬：ボールと自ゴールを結ぶ直線上に入った時
- キーパー報酬：自ゴールに近づいた (2m 以内) 時

実験 1(DR-FASSFPS の性能比較実験)

3 章で提案した「FASFPS」、「FASSFPS」、「DR-FASSFPS」の性能を比較するために図 3 の環境でシミュレーション実験を行った。初期状態は自ロボット 1 台、敵ロボット 5 台でありボールを捕獲し画面に向かって右側の青ゴールにシュート行う。敵ロボットはランダムな行動を取るものとした。

実験 2(戦略切替実験)

図 5 の初期配置においてロボットがどのような行動を学習するかの検証を行った。自ロボットと敵ロボットはそれぞれ 1 台ずつ配置し、敵ロボットは初期位置に固定した状態で自ロボットのポジション獲得を行った。自ロボットは初期配置から攻撃行動ならばボールを捕獲し、敵ゴールを目指してシュート行動を行い、守備行動ならばボールと敵ロボットと自ゴールの位置関係によりインターセプトやキーパーの守備位置を目指す。

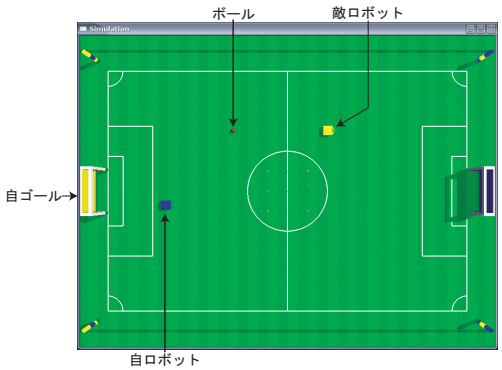

図 5: 初期状態 (実験 2)

5·1 実験結果および考察

実験は自ロボットがボール追跡や障害物回避などの基本行動 1 回を 1 ステップとし、自ロボットがボールを捕獲しシュート行動を行うという一連の行動を 1 試行とした。試行回数は実験 1 と実験 2 のそれぞれ 200 試行の行動獲得学習を行い累乗近似法を用いて近似を行い、10 試行毎の最大値と最小値の偏差をプロットした。

実験 1、2 の結果を図 6、7 に示す。実験 1 では調教者の負担を減らすために Shaping の自動化を行い、その結果と調教者によって手動で与えたものとの比較を行った。その結果、Shaping を自動化した手法 (DR-FASSFPS) は調教者によって適宜報酬を与える FASSFPS とあまり変わらない性能を示した。さらに分化強化を行い段階的な学習を行うことによって偏差のバラつきが少なくなることもわかった。これは Shaping 報酬を調教者が与える場合はその時その時で報酬を与えるタイミングが異なるのに対し、Shaping 報酬を自動で与えた場合はあらかじめ決められた条件で報酬が発生するために毎試行で安定した Shaping 報酬が得られる。そのため偏差のバラつきが少なくなったのではないかと考えられる。さらに分化強化を行って重点を置いた Shaping 報酬を与えることによって学習の初期から偏差の少ない効率的な学習が行えたのではないかと考えられる。

実験 2 は戦略切替ファジィルールを用いたポジション獲得手法 (Weight FPS) とボールが敵ロボットよりも近かったら攻撃行動、遠かったら守備行動と単純に切り替えるだけの FPS(Switch FPS) との比較を行ったが、図 7、8 の結果に示されるように守備行動のキーパーのポジショニングを学習した。学習結果から提案手法である Weiht FPS が最も良い性能を示した。これは実験 2 のような環境では、ボールの距離だけを用いて切り替えを行う Switch FPS では学習初期で攻撃と守備が激しく切り替わってしまうため学習が効率的に行われなかったと思われる。その点、重み付けを行って攻守を学習させた Weight FPS では、戦略切替ファジィルールを用いて報酬を度合いに応じて与えているため比較実験よりも報酬量が減っているにも関わらず良い結果となったと考えられる。

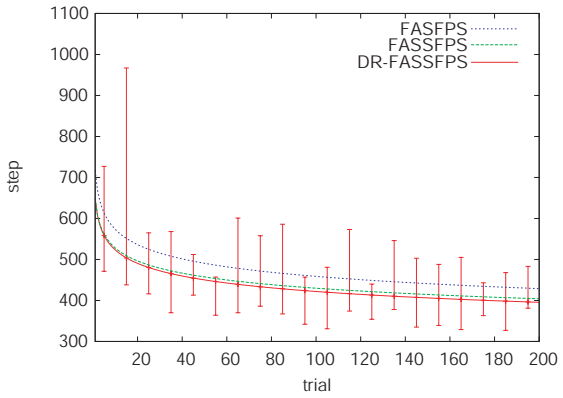

図 6: 実験 1：各試行におけるステップ数の推移 (偏差：DR-FASSFPS)

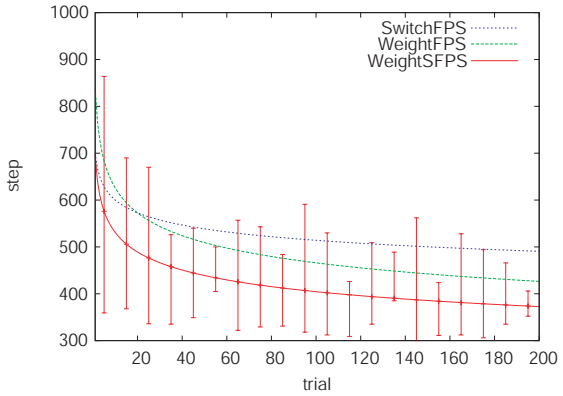

図 7: 実験 2：各試行におけるステップ数の推移 (偏差：Weight FPS)

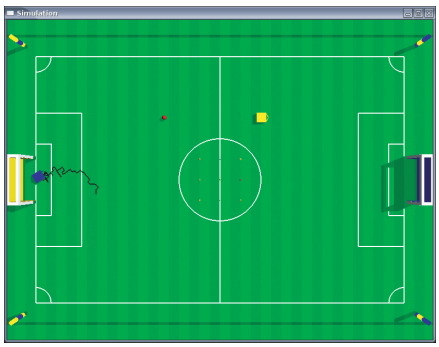

図 8: 実験 2：自ロボットが辿った軌跡

6. 結言

本研究では FPS に Shaping の概念を導入した手法 (SFPS)、ファジィルールを用いて Profit Sharing のパラメータのチューニングを行った手法 (FASFPS)、Shaping の自動化を行い分化強化の概念を導入した手法 (DR-FASSFPS) を提案した。シミュレーション実験により DR-FASSFPS は全体の報酬量が減少したにも関わらず、FASSFPS と同等以上の性能を示すことが確認できた。

参考文献

[1] 小林重信, 木村元, 小野功, “生物的適応システム—進化・学習のアルゴリズムと創発システム論—,” 計測と制御, Vol.40, No.10, pp.752-757 (2001)

[2] A.Ueno, H.Takeda, T.Nishida, “Learning of the way of abstraction in real robots,” *Proc. of 1999 IEEE International Conf. on Systems, Man, and Cybernetics SMC99*, Vol.1, No.10, pp.II746-II751 (1999)

[3] 近藤敏之, 伊藤宏司, “進化的適応戦略を用いた強化学習法,” 第 *14* 回自律分散システム・シンポジウム資料, pp.1-6 (2002)

[4] 田渕一真, 谷口忠大, 椹木哲夫, “模倣学習と強化学習の調和による効率的行動獲得,” *The 20th Annual Conference of the Japanese Society for Artificial Intelligence*, pp.212-215 (2005)

[5] 花香敏, 前田陽一郎, “自律移動ロボットの戦略獲得のためのファジィ状態分割型強化学習,” 第 *23* 回日本ロボット学会学術講演会, CD-ROM, 3E12 (2005)

[6] 前田陽一郎, 花香敏, “Shaping 強化学習を用いた自律エージェントの行動獲得支援手法,” 日本知能情報ファジィ学会誌, Vol.21, No.5, pp.722-733 (2009)