B2-2　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

福井大学 工学部　知能システム工学科
花香　敏　(指導教官：前田 陽一郎 助教授)

1　緒 言

近年、複雑な環境下で自律移動ロボットが適応的行動を獲得するための手法が盛んに研究されている。従来の手法では環境が動的になると制御ルールのチューニングが困難であったり、学習時間が非現実的になる、などの問題があった。そこで、最近ではこのような問題を解決するため様々な改良案が提案されている [1][2]。しかしながら、これらの手法では動的な環境で動作するロボットのためのより高次元な行動戦略を短時間で取得することは困難である。

そこで本研究では、当研究室で既に提案されている複雑なタスクをサブタスクに分割した行動決定ファジィルールを作成し、上位の行動選択ファジィルールで統合する階層型ファジィ行動制御 [3] を基に、上位の行動選択部を状態空間に連続値を扱うことのできるファジィ状態分割型 Profit Sharing 法 (FPS)[4] を用いて行動戦略を獲得する手法を提案する。さらに、学習時間の短縮を行なうために、学習に用いる入力状態はセンサなどから得られる環境状態を直接入力とするのではなく、ファジィ推論を用いてマクロな状態認識を前段で行ない、状態の次元数を減少させる手法も提案する。本論文では、提案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作成し、従来の Profit Sharing 法や、状態次元数を減少させない FPS との比較を行ったので、その結果についても報告する。

2　戦略獲得が可能な階層型ファジィ行動制御

図 1 に本手法で提案する階層型ファジィ行動制御手法の概念図を示す。本手法ではまず、複雑なタスクをサブタスクに分け、個々の独立したタスクに対し、分割したタスク数だけ行動決定ファジィルールを作成する。サブタスクに分けることによって、それぞれのルールのチューニングは比較的容易に行なうことができる。さらに、サブタスクを状況によって上位の行動選択部で切り替えるが、ここでは、行動選択部に FPS を用いることによって行動戦略の自動獲得を行なうことを考える。FPS を用いることで、連続的な状態空間を滑らかに扱うことができ、Q-Learning などの環境同定型の学習手法に比べると、環境にマルコフ性を仮定していないので実環境でも適応可能な手法となっている。

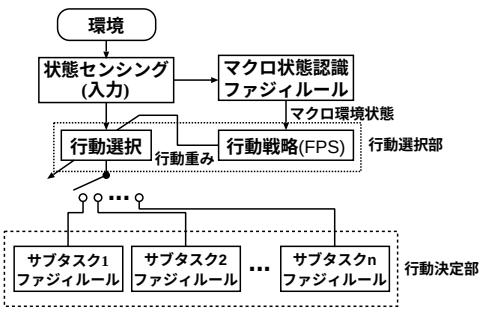

図 1. 階層型ファジィ行動制御手法の概念図

さらに、本手法では学習時間の短縮のため、学習に用いる入力状態をセンサなどから得られる環境状態を直接学習器の入力状態に用いるのではなく、前段でファジィルール (マクロ状態認識ファジィルールと呼ぶ) を用いてマクロな認識を行ない、状態の次元数を減少させることを試みた。学習者は環境から n 次元連続ベクトル $S = s_1, ..., s_n$ を観測した場合、一例として、この中から関連性のある情報を複数選んでマクロ状態認識を行なったと仮定すると、例えばファジィルールは以下のようになる。

マクロ状態認識 $1 : IF\ s_1\ is\ A_y\ and\ s_2\ is\ B_y\ THEN\ S_{1z}$

マクロ状態認識 $2 : IF\ s_3\ is\ C_y\ and\ s_4\ is\ D_y\ THEN\ S_{2z}$

マクロ状態認識 $3 : IF\ s_2\ is\ B_y\ and\ s_5\ is\ E_y\ and\ s_6\ is\ F_y\ THEN\ S_{3z}$

ここで、$s_x(x = 1, ..., n)$ はセンサより入力された状態、A_y,...,$F_y(y = 1, ..., l;\ y$ は状態分割数) は分割された状態を示すメンバーシップ関数、$S_{1z}, ..., S_{3z}$ はマクロ環境状態 ($z = 1, ..., l$; z はマクロ状態分割数) を示す。このように観測された情報を用いてマクロ状態認識ファジィ推論を行ない、入力状態の次元数を減少させる。

PS 法では状態と行動の組をルール系列として記憶し、その重み付け (ここではこの重みのことを行動重み w_{ik} と呼ぶ) を基にルーレット選択により行動選択を行なう。ここで、入力状態を n 次元として、それぞれの状態をすべて等しい分割数 l とし、取り得る行動が m 通りあるとすると式 (1) によって、行動重みの数が決まる。例えば、上のファジィルールの場合、状態次元数 n が 6 次元で状態分割数 l を 3 とし、取り得る行動 m が 3 と仮定すると行動重み w_{ik} は 2187 個になり、マクロ環境状態は状態次元数が 1/2 の 3 次元になっているので、行動重みは 87 個となり、1/27 に減少する。このように、状態次元数が減少すると、更新すべき行動重みの数が激減するため、学習時間の短縮につながる。

$$w_{ik} = l^n \times m \tag{1}$$

マクロな認識によって環境からの入力状態を p 次元まで減少させることできたとすると、p 次元連続ベクトル $D = d_1, ..., d_p$ を入力状態として、以下のようなファジィルールにより、サブタスクの各行動重み w_{ik} を導出する。

$$R_1 : IF\ d_1\ is\ B_{11}\ and ... d_p\ is\ B_{1n}\ THEN\ w_{11}, w_{12}, ..., w_{1m}$$
$$\vdots$$
$$R_q : IF\ d_1\ is\ B_{q1}\ and ... d_p\ is\ B_{qp}\ THEN\ w_{q1}, w_{q2}, ..., w_{qm}$$

ここで、$B_{ij}(i = 1, ..., q; j = 1, ..., p)$ は状態を表すファジィ集合でファジィルールの数は q 個、マクロ環境状態の数は p 個であることを示す。また各ファジィルールの後件部シングルトンは行動重み $w_{ik}(k = 1, ..., m)$ であり、m はサブタスクの数を示している。式 (2) により、前件部の適応度 C_i を計算し、式 (3) の重み付け平均により最終的な推論結果である w^*_{ik} が導出される。提案手法では w^*_{ik} は下位のサブタスクの行動重みとなる。行動選択はこの w^*_{ik} を用いて、ルーレット選択を行ない、試行錯誤的な探索を行なう。

$$C_i = B_{i1}(d_1) \wedge B_{i2}(d_2) \wedge ... \wedge B_{in}(d_n) \tag{2}$$

$$w^*_k = \frac{\sum_{i=1}^{q} C_i \cdot w_{ik}}{\sum_{i=1}^{q} C_i} \tag{3}$$

学習初期はすべての w^*_{ik} に一定の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、エピソードにその履歴を記憶する。状態遷移は全ルールの前件部の適合度で最も大きな値をとったルールを現在の状態を示すものと考え、行動は状態遷移を行なった直後の行動が状態を変化させた時の行動と考え、この 2 つの行動と状態の組み合わせをエピソード

記憶に蓄えていく。そして、報酬 r が得られた場合, 報酬を与えられた時点を 0 ステップ考え、h ステップ前で行動 i を取った場合の行動重みは式 (4),(5) に従い、更新される。式 (4) は等比減少関数を示しており、α は学習率、γ は割引率、h は報酬が得られるまでのステップ数を示す。

$$f(h) = \gamma^h \cdot r \tag{4}$$

$$w_{ik}(d) = w_{ik}(d) + \alpha f_i(h) \tag{5}$$

3 サッカーシミュレーション

本手法の有効性を示すためにサッカーシミュレータを作成した。ここでは、シミュレーション条件および学習に用いた各パラメータの設定を中心に報告する。

3.1 サッカーシミュレータの製作

本研究で作成したシミュレータは Linux 上で Xlib のグラフィックスライブラリおよび $Motif$ ウィジェットを利用し、C 言語を使用して開発した。シミュレータの画面を図 2 に示す。

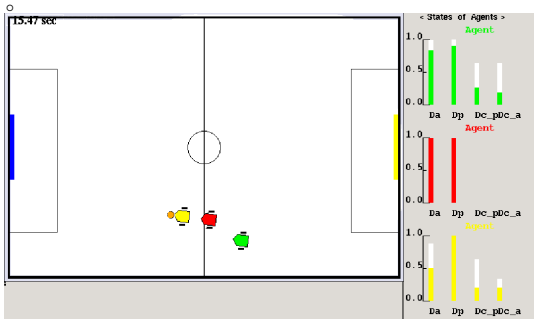

図 2. シミュレーションの画面構成

問題設定として実際のサッカーの練習で一般に行なわれている 2 対 1 のボールキープトレーニングを対象とした。学習者は 2 台で 1 チームを組み、ボールだけをひたすら追いつづける敵ロボットにボールを取られないように、また、敵がボールを持っているときはボールを奪い取るようにフィールド上を動き回る。ロボットはロボカップの中型リーグ規格のロボットを想定して、左右独立駆動方式の機構をもち、本研究室で提案されているマルチ全方位ビジョンシステム (MOVIS)[3] を搭載しているものと仮定した。MOVIS は全周における物体の距離および方位情報を比較的精度よく一度に取得できる。

3.2 シミュレーションの設定

観測できる情報としてはボールの相対距離・方位、障害物の相対距離・方位 (他のロボットとタッチラインを障害物とみなす)、味方ロボットの相対距離・方位・絶対位置、敵ロボットの相対距離・方位・絶対位置、および自分自身の絶対位置である。ロボットの取り得る基本行動としては以下のものを想定した。

- 障害物回避
- ボール追跡
- ボール出し (パス)
- ボール受け (レシーブ)

それぞれの行動は簡略化ファジィ推論を用いた行動決定ファジィルールにより行動制御される。さらにこれらの行動選択は、表 1 に示すような観測された環境情報を基にマクロ状態認識ファジィルールで一旦マクロ環境状態に落とし、これを入力として FPS により行動選択の際に用いる行動重みを学習する。

Da はロボットがどれだけ危険な状況にいるかという度合いを表し、Dp はいかにボールを追いかけようとしているかという度合いを表し、Dca は味方にパスをするといった自分から仲間に助けを働きかける行為ができる状態にいるかという度合いを表し、Dcp はパスを受け取るといった仲間からの協調行動を受け入れる度合いを表す。シミュレータ上ではこのパラメータが各ロボットごとに確認できるように図 2 に示す画面右のバーグラフでリアルタイムに表示される。ここで導出された値から式 (2)、(3) を用いて各行動重みを導出し、これを基にルーレット選択により行動選択が行なわれる。そして、報酬が得られた場合は式 (4),(5) に従い、各行動重みに分配され、次第に、最適な行動戦略を獲得できるようになる。尚、紙面の関係上、具体的なファジィルールおよびシミュレーション結果については卒研発表会にて報告する。

表 1. マクロ状態認識に用いた環境情報

観測された環境情報	マクロ環境状態
障害物の距離・方位	回避度 (Da)
ボールの距離・方位	追跡度 (Dp)
ボールの距離・方位 味方ロボットの距離・方位 敵ロボットの方位	能動協調度 (Dca)
ボールの距離・方位 味方ロボットの距離・方位 敵ロボットの方位	受動協調度 (Dcp)

4 実機ロボットを用いた実験

シミュレータ環境下で獲得した戦略が実環境下でも有効性があることを検証するため、実機ロボットに獲得した戦略を導入し、フィールド実験を行なう予定である。図 3, 図 4 に本年度、当研究室で試作したロボカップの中型リーグ規格のサッカーロボットを示す。ロボットはシミュレータ上のロボットと同様に左右独立駆動方式の機構をもち、最大速度 $2m/s$ で走行することが可能である。

図 3. ロボットの外観

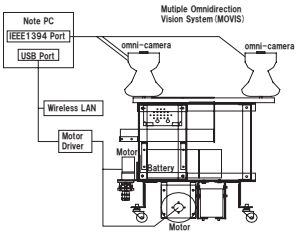

図 4. システム構成図

5 結 言

本研究では、自律移動ロボットのための FPS を用いた戦略獲得が可能な階層型ファジィ行動制御手法を提案した。ロボットの基本的な行動は、あらかじめファジィルールを用いて与えているのでより高度な行動戦略を獲得する部分に強化学習を適用した。これにより、従来手法より比較的短時間で戦略レベルのより高度な行動戦略を獲得できることがわかった。今回はマクロな状態認識によりロボットの学習時間を短縮することができたが、人間の認知機構などを考慮して今後改良を行なっていく必要があると考える。また、今回は割引率、エピソードの長さ、強化関数などはすべて固定して学習を行なったが、これらの要素は PS 法において、重要な役割を果すため、与え方を検討する必要があるものと思われる。

参考文献

[1] 港 隆志, 浅田 稔: "環境の変化に適応する移動ロボットの行動獲得," 日本ロボット学会誌, *Vol.18, No.5*, pp.706-712 (2000)

[2] 高橋 泰岳, 浅田 稔: "複数の学習器の階層的構築による行動獲得," 日本ロボット学会誌, *Vol.18, No.7*, pp.1040-1046 (2000)

[3] W.Shimizuhira, K.Fujii, Y.Maeda: "Fuzzy Behavior Control for Autonomous Mobile Robot in Dynamic Environment with Multiple Omnidirectional Vision System," *In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'04)*, pp.3412-3416 (2004)

[4] 前田 陽一郎, 牧田 研吾: "ファジィ状態分割型強化学習を用いた協調行動シミュレーション," 第 *11* 回日本ファジィ学会北信越支部シンポジウム講演論文集, pp.49-52 (2002)