A2-05 ファジィ状態分割型 Profit Sharing 法を用いた自律移動ロボットの行動学習

牧田 研吾 (指導教官： 前田 陽一郎 助教授)

福井大学 工学部 知能システム工学科

1 緒 言

近年、機械学習あるいはロボット工学の分野で研究が盛んなテーマとして強化学習法がある。強化学習法の代表的な手法として Q-Learning 法 (以下 QL 法) や Profit Sharing 法 (以下 PS 法) があげられる。

しかしながら、従来の強化学習法では主に状態、行動が離散的であるものが対象とされており、連続的に状態や行動が変化する実世界の対象は扱いにくい。実際に現実の複雑な環境での学習を必要とする際には、連続値を含めた多様な入出力データを取り扱う必要がある。そのため QL 法において、連続値の状態、行動を取り扱うことを可能にするための研究も進められている [1]。また、従来の強化学習では学習目標が多数ある場合、同時に複数の学習目標を満たすように学習を進めることも困難である。これに対し当研究室では、QL 法を用いてより複雑な環境下で複数の目的が競合するような場合においても学習ができ、かつ報酬量を学習者の状態によって適宜変化させ、連続値の状態を取り扱うことを可能にしたファジィ状態分割型 Q-Learning(以下 FQL) をすでに提案している [2]。

本研究では、FQL をさらに改良し、状態分割にファジィ推論を用い、強化学習法として PS 法を用いたファジィ状態分割型 Profit Sharing 法 (以下 FPS) を提案する [3]。PS 法は適用可能な環境のクラスが広く、連続的な行動の素早い学習が可能であるため、QL 法と比較すると非マルコフ決定過程へ適用しやすい。本手法を用いて内部エネルギーを持つエージェントがエネルギーを補給しつつ、移動物体を追跡する計算機シミュレーションを行ったのでその結果についても報告する。

2 ファジィ状態分割型 Profit Sharing 法

PS 法とは報酬を得たときにそれまでに使用された感覚入力 (状態) と行動の組であるルール系列を一括的にエピソード単位で強化する学習法である。エピソードとは初期状態あるいは報酬を得た直後から報酬までのルール系列のことである。PS 法ではルール系列に付加された行動重みを強化する。ここで行動重みとは、状態と行動の組み合わせで定義されたルール系列の重要性を示すものである。また行動重みを強化する強化関数として一般に等比減少関数を用いる。(図 1 参照) 尚、PS 法では学習過程における行動選択の方法としては、ルーレット選択が良い性能を示すことが経験的に知られている。そこで、ロボットが行動を選択する際にはルール系列に付加された行動重みのルーレット戦略で行動を決定する。

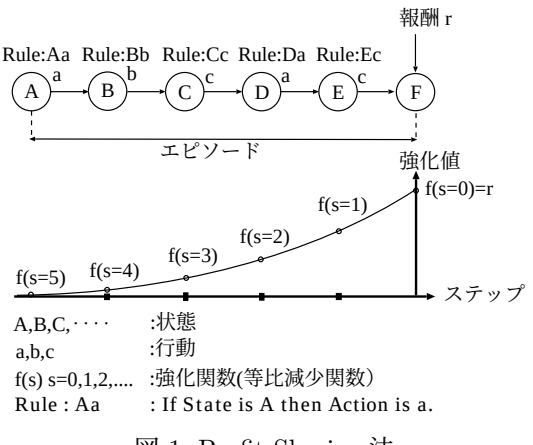


図 1. Profit Sharing 法

提案手法では、状態に連続値を取り扱えるようファジィ推論を導入するため、各変数をファジィ変数とし、メンバーシップ関数を用いて状態空間を表現する。学習者がある環境から状態としてn次元連続値ベクトル $x = (x_1, x_2, ..., x_n)$ を観測した時、PS 法の行動決定を行う各ファジィルールの行動の重み w_{ik} を導出するためのファジィルールは以下のようになる。

$$R_i \ : \ IF \ x_1 \ is \ A_{i1} \ and \ ... \ and \ x_n \ is \ A_{in} \ then \ w_{ik} \quad (1)$$

ここで $A_{ij}(i = 1, ..., l), (j = 1, ..., n)$ は状態を表すファジィ集合で、状態の数 (ファジィルールの数) は l 個存在する。また各ファジィルールの後件部シングルトン $w_{ik}(k = 1, ..., m)$ は PS 法におけるルール系列の行動重みを示しており、本手法においてはある状態における行動の重みを表し、行動の数 (m 個) だけ存在する。

式 (2) により、各ファジィルール R_i についての前件部の適合度 C_i を計算し、これらの重み付き平均により以下の式 (3) に従って最終的な推論結果である行動重み w_k^* が導出される。

$$C_i = A_{i1}(x_1) \wedge A_{i2}(x_2) \wedge ... \wedge A_{in}(x_n) \quad (2)$$

$$w_k^* = \frac{\sum_{i=1}^{l} C_i \cdot w_{ik}}{\sum_{i=1}^{l} C_i} \quad (3)$$

図 1 において、報酬 r を与えられたステップを 0 とし、そこからさかのぼって s 番目に発火したのが i 番目のファジィルールであったと仮定すると、そのときの行動重み w_{ik} は以下の式 (4),(5) に従って報酬値を分配し、更新される。

$$w_{ik} = w_{ik} + \alpha f(s) \quad (4)$$

$$f(s) = \gamma^s \cdot r \quad (5)$$

ここで α は学習率 ($0 \leq \alpha \leq 1$)、γ は割引率 ($0 \leq \gamma \leq 1$) である。式 (5) は等比減少関数であり、報酬を与えられた時点に近いファジィルールほど行動重みが大きく更新されることになる。尚、この場合のルールの発火順であるが、ここでは全ルールの前件部適合度 (min 演算後の値) で最も大きな値をとったルールが現在の状態を示すルール (発火ルール) と考える。

3 物体追跡シミュレーション

今回作成した物体追跡シミュレータの概要を説明する。図 2 の物体追跡シミュレータで用いた前提条件は次のとおりである。1) 各エージェント、餌場、捕獲物体の初期位置はランダムに配置される。2) エージェントは餌場にたどり着くことにより自らのエネルギーを補給しつつ、物体を追跡し捕獲する。3) 餌場の保有量は有限であり、エネルギーが無くなった餌場は消滅し、また新たな餌場がランダムに出現する。4) 自己以外のエージェントは障害物とみなす。5) 移動物体は追跡に来たエージェントが近づいてきた場合のみ遠ざかる方向へ逃げるものとする。

ここで各エージェントの行動ルールについて説明する。本シミュレーションでは、各エージェントは移動物体の捕獲、餌場でのエネルギー補給、障害物の回避を学習目標として学習を行う。各学習エージェントの取り得る行動としては、1) 移動物体の追跡に向かう 2) 餌場に向かう 3) 障害物を回避するの 3 行動を取れるものとする。

エージェントは自己の前方 8 方向に 15 度間隔でセンサを持っているものと仮定する。このセンサによって障害物の距離と方位を測定し、それを入力としてファジィ推論により状態を認識する。

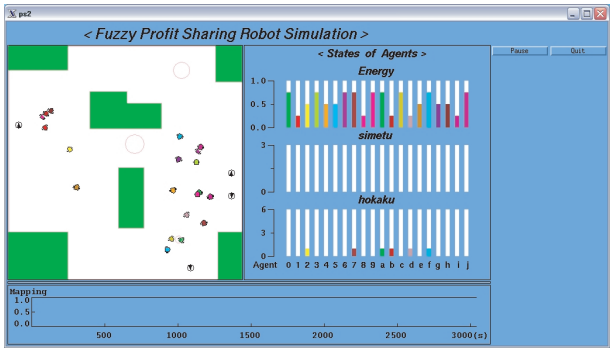
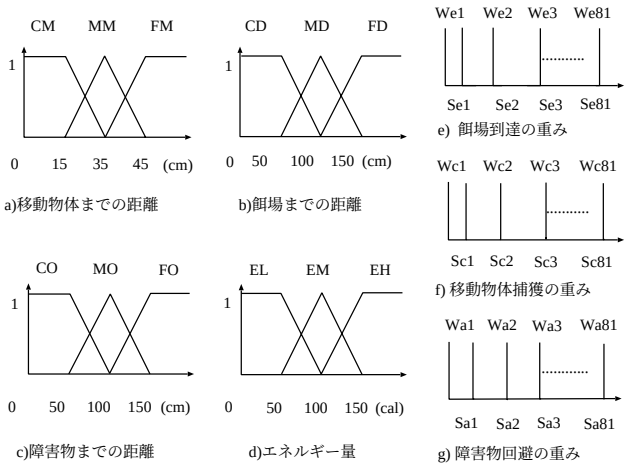


図 2. シミュレータの画面構成

本シュミレーションでは移動物体までの距離、餌場までの距離、障害物までの距離、エネルギー量を各エージェントの状態量 (前件部) として、3つの行動に対する行動重みを後件部として、図 3 のように設定した。



a)前件部メンバシップ関数　　**b)**後件部シングルトン

図 3. 本手法のメンバーシップ関数とシングルトン

以上のメンバーシップ関数より本シミュレーションで用いたファジィルールを図 4 に示す。本手法では後件部シングルトンが行動重み w_{ik} になっているため、エージェントの学習が進むにつれ、ルールの後件部シングルトンは報酬を与えられれば大きく、罰を与えられれば小さくなっていく。

4　シミュレーション結果および考察

本シミュレーションでは本手法の有効性を検証するため、従来の PS 法との比較を行った。PS 法でのエージェントおよび、移動物体の行動ルール等は本手法のシミュレーション条件と同等のものを使用した。

本シミュレータでは 20 体のエージェントまでシミュレーションが可能であるが、今回は 1 体だけでシミュレーションした結果を報告する。これはタスクを簡単にすることでより学習の結果を明確にするためである。また捕獲された移動物体は消えてしまうので一定時間ごとに復活させる。

シミュレーションの結果として、捕獲した移動物体の数 (捕獲数) とエネルギーが無くなり死滅した回数 (死滅数) の累積値で比較を行った。障害物との距離が危険領域内に入った回数 (衝突数) は今回参照しない。

その結果、捕獲数、死滅数ともに本手法が PS を上回る結果が得られた。50000 試行付近から FPS では死滅数がほぼ一定で増加していないのに対し、PS は 150000 試行でも増加し続けている。また FPS は死滅数を抑えながらも PS 法よりも多くの移動物体を捕獲している。よって FPS は効率的

エネルギー＼移動物体＼餌場・障害物		CM			MM			FM		
		CD	MD	FD	CD	MD	FD	CD	MD	FD
EL	CO	Wc1 We1 Wa1	Wc2 We2 Wa2	Wc3 We3 Wa3	・・・・	・・・・	・・・・	Wc7 We7 Wa7	Wc8 We8 Wa8	Wc9 We9 Wa9
	MO	Wc10 We10 Wa10	⋱							Wc18 We18 Wa18
	FO	Wc19 We19 Wa19		⋱						Wc27 We27 Wa27
EM	CO	⋮			⋱					⋮
	MO	⋮				⋱				⋮
	FO	⋮					⋱			⋮
EH	CO	Wc55 We55 Wa55						⋱		Wc63 We63 Wa63
	MO	Wc64 We64 Wa64							⋱	Wc72 We72 Wa72
	FO	Wc73 We73 Wa73	Wc74 We74 Wa74	Wc75 We75 Wa75	・・・・	・・・・	・・・・	Wc79 We79 Wa79	Wc80 We80 Wa80	Wc81 We81 Wa81

図 4. 本手法のファジィルール

な捕獲と死滅をしないという二つの目標をバランス良く学習できていると考えられる。

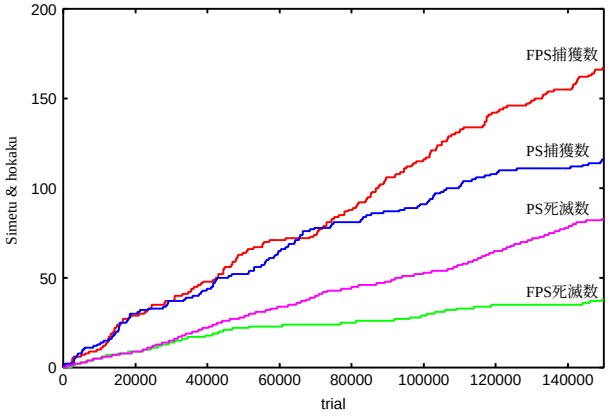


図 5. シミュレーション結果

5　結 言

本研究では、従来の強化学習法では取り扱うことが出来なかった連続値を取り扱うため、状態分割にファジィ推論を用いたファジィ状態分割型 Profit Sharing 法（FPS）を提案した。

さらに本手法を検証するため、物体追跡シミュレーションを行い、PS 法との比較を行った。本手法が良い結果を得た要因として、ファジィ推論により状態を連続値として取り扱うことができ, さらに 2 つの異なる学習目標に対して柔軟な学習ができたことが考えられる。

複数の学習エージェントが存在するマルチエージェント系や、実機ロボットへの搭載が今後の課題として残っている。

参考文献

[1] 堀内 匡, 藤野 昭典, 片井 修, 椹木 哲夫: “経験強化を考慮した Q-Learning の提案とその応用,” 計測自動制御学会論文集, Vol.35,No.5, pp.645-653(1999)

[2] 前田 陽一郎: “ファジィ状態分割型 Q-Learning を用いたマルチエージェントシミュレーション,” 第 *17* 回ファジィシステムシンポジウム, pp.419-422 (2001)

[3] 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シンポジウム, pp.49-52 (2002)