A2-1 個体および集団貢献度評価に基づくマルチエージェント強化学習の報酬分配に関する研究

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
張 坤 (指導教員：前田 陽一郎)

1. 緒 言

強化学習はシングルエージェントを対象に開発された手法であり、マルチエージェントで強化学習をそのままを適用すると、複数のエージェントが協力して報酬を得ることができたときに各エージェントの行動にどのように報酬を分配するかという報酬分配問題が発生する [1]。個々のエージェントの目標はそれぞれの期待獲得報酬を上げることであるが、報酬分配を適切に行わなければ各エージェントの行動に悪影響を与えるだけでなく、システム全体の能力低下にもつながる。

宮崎ら [2] は期待獲得報酬が正となるための各エージェントへの報酬分配に関する必要十分条件を出したが、この手法では報酬に貢献していないエージェントにも報酬が与えられる可能性がある。また保知ら [3] は貢献度に基づき報酬分配する手法を提案したが、直接的に貢献したエージェントと間接的に貢献したエージェントが得る報酬が共に一定であるという問題が残されている。本研究室の池田らは報酬を発生する必要条件により、その条件をクリアするまでに要した時間に基づく貢献度による報酬分配手法を提案したが、貢献度を条件をクリアした時間のみにより決定している点でやや問題がある。

そこで本研究では、マルチエージェント強化学習で起こる報酬分配問題において、個体行動評価とファジィルールを用いる協調行動評価により、個々のエージェントの全ての行動を評価し、個体貢献度とシステム全体に対する協調貢献度を求めて、個々のエージェントに報酬を分配するマルチエージェント Profit Sharing(MAPS) 手法を提案する [4]。提案手法の有効性を検証するため、獲物追跡問題を例題にシミュレーション実験を行ったので、その結果についても報告する。

2. 貢献度による報酬分配法

マルチエージェント強化学習ではシングルエージェント強化学習の問題点と共に、なおかつマルチエージェント強化学習独特の問題点が起こる。個体の行動をシステム全体で考えた場合、個体が同じ環境で同じ行動をしてもその時間やシステム全体の状況によっては評価は変わるためである。

そこで報酬発生までに各エージェントの個々の行動に対する評価を通じて、報酬発生時に個々の「貢献度」を求めて、その貢献度を基としてエージェントに報酬を与える手法を提案する。すなわち、貢献度を設けることで個々の行動に基づきシステム全体の評価をし、個体の行動をシステム全体に生かすことを考える。

2·1 行動評価について

本手法の行動評価では自身の利益を最大化するための個体行動評価と、自身の利益を考慮しないで全体の利益を最大化するための協調行動評価により構成される。

2·1.1 個体行動評価手法

複数エージェントが存在する場合には、状態遷移は自分の行動だけでなく、他のエージェントとの協調行動になることが多い。しかし、個体行動評価では相互的な影響を考慮せず、他のエージェントの行動を環境の一部とし、シングルエージェントのように自分自身の報酬を最大化することを目的とする。

本研究での個体行動評価手法は集団追跡問題またはグリッド探索問題に適応でき、目標を達成すると、仮に最高評価値 100 を得るものとする。そこで、個体行動評価を式 (1) のように定義する。

$$E_i(t) = \frac{100}{d_t} \tag{1}$$

ここで、
t: 時刻
$E_i(t)$: 個体 i の時刻 t の個体行動評価値 (達成すると、最高評価値 100 を得る)
d_t: 時刻 t における目標までの距離 (d_t>0)

2·1.2 協調行動評価手法

マルチエージェントシステムでは相互に協調しない限り、あるタスクを達成することができない。協調行動評価では自分自身の行動のみを考慮するのではなく、全体のエージェントの協調行動を考慮し、全体の報酬の最大化を目的とする。

マルチエージェントの協調行動を合理的に評価するため、集団追跡問題における協調行動評価ではファジィルールにより評価することを考える。ファジィルールへの入力は集団分散度 D と目標遠隔度 F である。集団分散度 D とは、すべてのエージェントの集団の中心までの平均距離を示し、目標遠隔度 F とは、集団の中心から目標までの距離を示す。ファジィルールの出力は協調行動評価値 E_g である。前件部メンバーシップ関数を図 1 と図 2 に、後件部シングルトンを図 3 に、ファジィルールを表 1 に示す。

集団の協調行動評価

集団の協調行動評価とは全体のエージェントがファジィルールにより、すべてのエージェントの協調行動を評価するメカニズムである。時刻 t の n 台エージェントの協調行動評価値 $E_g^n(t)$ は集団分散度 D と目標遠隔度 F に基づき、表 1 のファジィルールにより簡略化ファジィ推論の重心合成法で求められる。

個体の協調行動評価

個体の協調行動評価ではファジィルールにより個体の集団行動を評価するが、集団の協調行動評価と比べると、個体の協調行動評価では個々のエージェントは

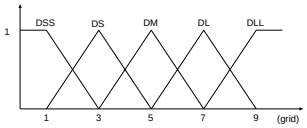

図 1: 集団分散度 D

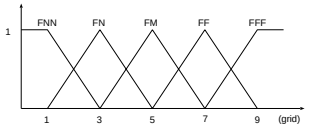

図 2: 目標遠隔度 F

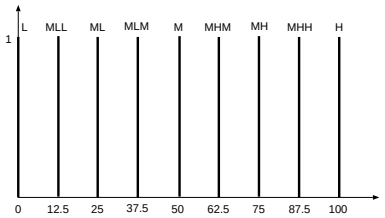

図 3: 集団評価値 E_g

表 1: ファジィルール

F ＼ D	DSS	DS	DM	DL	DLL
FNN	H	MHH	MH	MHM	M
FN	MHH	MH	MHM	M	MLM
FM	MH	MHM	M	MLM	ML
FF	MHM	M	MLM	ML	MLL
FFF	M	MLM	ML	MLL	L

個体行動に対する唯一の協調行動評価値をもち、全体と協調するかどうかを判断できるようになる。個体の協調行動評価値は式 (2) により定義される。

$$E_g^i(t) = E_g^n(t) - E_g^{n-1}(t) \tag{2}$$

ここで、
t: 時刻
n: エージェントの総数
$E_g^i(t)$: 時刻 t の個体 i の協調行動評価値
$E_g^n(t)$: 時刻 t の n 台のエージェントの協調行動評価値
$E_g^{n-1}(t)$: 時刻 t の自分自身以外の $(n-1)$ 台エージェントの協調行動評価値

個体 i の協調行動評価値を求めるため、個体 i を含める n 台のエージェントの協調行動評価値 $E_g^n(t)$ と個体 i を含めない $(n-1)$ 台のエージェントの協調行動評価値 $E_g^{n-1}(t)$ をそれぞれ協調行動評価値のファジィ推論により計算し、個体 i がいる場合といない場合の差を通じて、個体 i の協調行動評価値を定義する。

2·2 貢献度について

貢献度とはエージェントが目標を捕捉する (目標達成) までにそれまでの行動が全体に貢献した度合いを示す。具体的には、すべての時刻に対する行動を評価し、その評価値に基づいて、あるエピソードに対する貢献度を計算する。行動評価値 E は前述の個体行動評価値、集団の協調行動評価値、個体の協調行動評価値の三つで構成されている。

2·2.1 個体貢献度

個体貢献度とはエージェント個体は相互的な協調行動を考慮せず、自分自身の報酬を最大化することを目的とする個体行動評価により計算される貢献度である。個体貢献度 C_i は、式 (3) のように定義される。

$$C_i = \sum_{j=0}^{a} \gamma^j E_i(t_0 - j) \tag{3}$$

ここで、
C_i: 個体 i の個体貢献度
t_0: タスク終了時刻
$E_i(t_0 - j)$: 時刻 $(t_0 - j)$ の個体 i の行動評価値
γ: 割引率 $(0 \leq \gamma \leq 1)$
a: 貢献度に関与した有効行動時間

2·2.2 集団の協調貢献度

集団の協調貢献度とは自分自身の行動は考慮せず、集団エージェントの協調行動を考慮し、全体の報酬を最大化することを目的とする集団の協調行動評価により計算される貢献度である。集団の協調貢献度 C_g は、式 (4) のように定義される。

$$C_g = \sum_{j=0}^{a} \gamma^j E_g^n(t_0 - j) \tag{4}$$

ここで、
C_g: 全エージェントの協調行動評価値による集団の協調貢献度
$E_g^n(t_0 - j)$: 時刻 $(t_0 - j)$ のエージェント集団の協調行動評価値

2·2.3 個体の協調貢献度

個体の協調貢献度とは個々のエージェントは集団全体で唯一の協調行動評価値をもち、全体とどの程度協調しているかを評価する集団の協調行動評価値により計算される貢献度である。個体の協調貢献度は、式 (5) のように定義される。

$$C_g^i = \sum_{j=0}^{a} \gamma^j E_g^i(t_0 - j) \tag{5}$$

ここで、
C_g^i: 個体 i の協調行動評価値による個体の協調貢献度
$E_g^i(t_0 - j)$: 時刻 $(t_0 - j)$ の個体 i の協調行動評価値

2·3 貢献度に対する報酬分配手法

マルチエージェント強化学習において、システムに与えられる基本報酬を R とする。前節における貢献度の定義に基づき、これに対する報酬分配手法について本研究では比較のため以下の 4 種類の分配方法を提案する。

2·3.1 個体貢献度に対する報酬分配手法 1

ここでは基本報酬をすべてのエージェントの個体貢献度により分配し、個々のエージェントが得る報酬を式 (6) により計算する。

$$r_i = R \cdot \frac{C_i}{\sum_{j=1}^{n} C_j} \tag{6}$$

ここで、
r_i: 個体 i の貢献度に対する報酬
R: 基本報酬
n: エージェントの総数

2·3.2 個体の協調貢献度に対する報酬分配手法 2

ここでは個体自身を含めた全体 (n 台) の協調行動評価と自身を含めない (n-1) 台の協調行動評価の差を通じて、得られた個体の協調貢献度に基づいて、報酬を式 (7) により分配する。

$$r_i = R \cdot \frac{C_g^i}{\sum_{j=1}^{n} C_g^j} \tag{7}$$

2·3.3 個体貢献度と集団の協調貢献度に対する報酬分配手法 3

ここでは個体貢献度に対する報酬分配手法 1 に、集団の協調貢献度を加え、個体と全体の総合貢献を考慮して報酬を式 (8) により分配する。

$$r_i = R \cdot \frac{C_i + C_g}{\sum_{j=1}^{n} C_j + C_g} \tag{8}$$

2·3.4 個体貢献度と個体の協調貢献度に対する報酬分配手法 4

ここでは個体と全体の総合貢献を考慮するが、個体貢献度と集団の協調貢献度に対する報酬分配手法 3 と比べ、集団の一員としての集団の協調貢献度ではなく、個体自身の個体と集団協調に基づく貢献度より報酬を式 (9) により分配する。

$$r_i = R \cdot \frac{C_i + C_g^i}{\sum_{j=1}^{n} C_j + \sum_{j=1}^{n} C_g^j} \tag{9}$$

3. 平均必要達成時間に対する報酬分配手法

貢献度による報酬分配手法は本研究で扱っているような集団追跡問題などのタスクに対する貢献度合いが測れる問題には有効であるが、そうでない問題になると、行動評価値の再定義が必要で、汎用性がない。そこでより汎用的に利用できる手法として、ここではエピソード達成時間のみに着目した報酬分配手法を提案する。

3·1 平均必要達成時間の定義

本手法ではエピソードを完成するために必要となる平均的な達成時間の長さによって、報酬分配を行なうことを考える。平均必要達成時間とは現在のエピソードとその前のエピソードに対する平均的な時間で、学習過程に伴って、動的に計算される。平均必要達成時間は、式 (10) で定義される。

$$T_a = \frac{\sum_{j=0}^{a} \alpha^j T_{(e-j)}}{\sum_{j=0}^{a} \alpha^j} \tag{10}$$

ここで、

T_a: 現在のエピソード e を完成するための平均必要達成時間
e: 学習したエピソード番号 (報酬を得た回数)
$T_{(e-j)}$: エピソード $(e-j)$ を完成するためにかかった実際の達成時間
α: 達成時間の割引率 $(0 \leq \alpha \leq 1)$
a: 平均必要達成時間 T に関与した有効エピソード数

ここでは偶然性を防止するため、式 (10) のように、現在のエピソード e を完成するための平均必要達成時間を報酬分配の基準とし、割引率と以前の実際の達成時間を通じて、現在の必要達成時間の合理性を維持して、随時更新する。

3·2 平均必要達成時間に対する報酬分配手法 5

エピソード e で、実際の達成時間 T_e は前節の平均必要達成時間 T_a より少ない時間の場合は多くの報酬をもらい、逆に多くの時間がかかる場合は少ない報酬をもらうという方針に基づき、報酬をもらえるエージェントは式 (11) に従って報酬を分配される。

$$r_e = R + R(\frac{T_a - T_e}{T_a}) \tag{11}$$

ここで、

r_e: エピソード e に対する報酬
T_e: エピソード e を完成するためにかかった実際の達成時間

4. シミュレーション実験

提案手法の有効性を示すために獲物追跡シミュレーション実験を行った。

4·1 獲物追跡シミュレーション

本研究で用いたシミュレータは Linux 上で OpenGL グラフィックス・ライブラリ及びその補助ライブラリである GLUT を用い、C 言語で製作した (図 4 参照)。

フィールドは 2 次元格子状で、ハンターエージェント (四角) は 4 体おり、行動、視野は全て等しいものとする。ハンターエージェント、獲物エージェント (三角) 共に上、下、左、右、左上、左下、右上、右下の 8 方向のいずれか 1 マスずつ動くことができる。獲物エージェントはハンターエージェントが視野に入ったとき、一番近いハンターエージェントから遠ざかる動きをする。

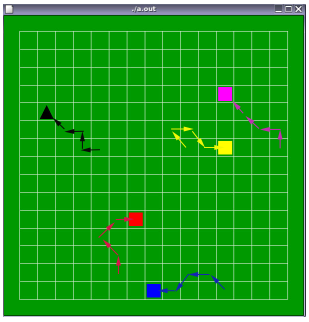

図 4: 作成したシミュレータ

4·2 シミュレーション結果および考察

実験ではハンターエージェントの視野は 5 × 5 とし、獲物エージェントの視野は十字型とした。実験では行動選択はルーレット選択を使用し、すべての手法とも 5 回ずつシミュレーションを行い、これらの 5 エピソードごとの平均と分散を表した結果を図 5〜図 9 に示した。縦軸は移動回数、すなわち達成までの試行回数を表す。また横軸は経験したエピソード数、すなわち学習回数を示す。

図 5 より、個体報酬のみに基づく報酬分配手法 1 は通常手法より悪くなった。これより、個々のエージェントの期待獲得報酬を最大化しても、システム全体の能力低下を示すことが分かった。図 6 より、個体の協調貢献度のみに基づく報酬分配手法 2 は通常手法とほとんど同じ性能を示した。この手法では個体と集団の協調いかんにかかわらず、全体の中心を尾行すれば高い報酬をもらえる。図 7 と図 8 より、個体と集団の利益を共に考慮したので、一方だけの利益を得ると報酬をもらい、両方の利益を同時に得るともっと高い報酬をもらうことができる。そのため、良い性能を示したものと考えられる。図 9 より、平均必要達成時間に対する報酬分配手法 5 は通常手法より良い結果を示したが、本手法は汎用性があるため追跡問題以外の問題にも適応できると考えられる。

また、ハンターエージェント視野が 5 × 5 から 7 × 7 になってもほぼ同様の結果を示したことから、異なる視野条件下においても有効性を確認することができた。

5. 結 言

本研究では個体貢献度、集団の協調貢献度、個体の協調貢献度の 3 つの貢献度と汎用的な平均必要達成時間を定義し、さまざまな組み合わせで 5 つのマルチエージェント Profit Sharing(MAPS) 手法を提案した。提案手法の有効性を検証するために獲物追跡問題を例題にシミュレーション実験を行った。その結果、個体貢献度と集団の協調貢献度に対する報酬分配手法が最も良好な結果を示した。さらに、異なる視野条件にパラメータを変更した実験でも同様の有効性を検証できた。

今後の課題として、複雑な環境や条件での適用、学習の高速化などが必要である。特に、追跡問題以外の応用に対する提案手法の改良が重要であると考える。また、エージェントが考慮すべき個体貢献度と協調貢献度の最適な割合を条件に応じてダイナミックに変動させる適応的なマルチエージェント強化学習手法への発展も考えられる。

参考文献

[1] 高玉 圭樹 著, マルチエージェント学習—相互作用の謎に迫る—, コロナ社 (2003)

[2] 宮崎 和光, 荒井 幸代, 小林 重信, “ Profit Sharing を用いたマルチエージェント強化学習における報酬分配の理論的考察, ”人工知能学会誌, Vol.14, No.6, pp.1156-1164 (1999)

[3] 保知 良暢, 松井 藤五郎, 犬塚 信博, 世木 博久, “ マルチエージェント強化学習における報酬発生条件に基づく貢献度判別と報酬分配, ”人工知能学会全国大会論文集, Vol.16th, 第 2 分冊, pp.2D3.02.1-2D3.02.4 (2002)

[4] 張 坤, 前田 陽一郎, “ 個体と集団の貢献度評価に基づくマルチエージェント強化学習, ”第 27回日本ロボット学会学術講演会, CD-ROM, RSJ2009AC1F1-03 (2009)

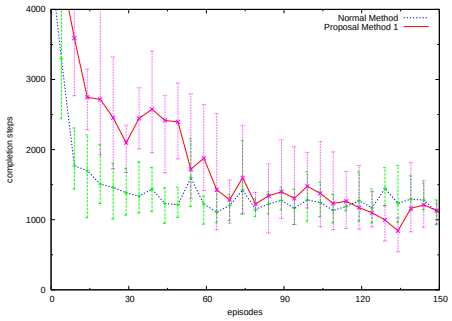

図 5: 通常の強化学習と報酬分配手法 1 の性能比較

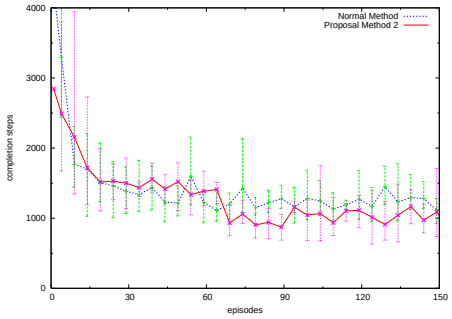

図 6: 通常の強化学習と報酬分配手法 2 の性能比較

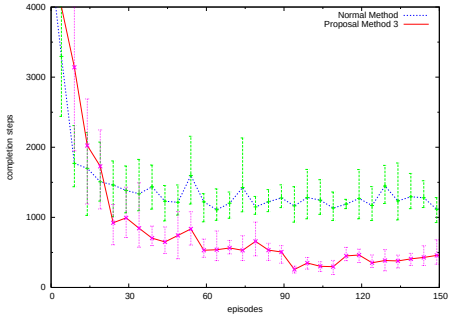

図 7: 通常の強化学習と報酬分配手法 3 の性能比較

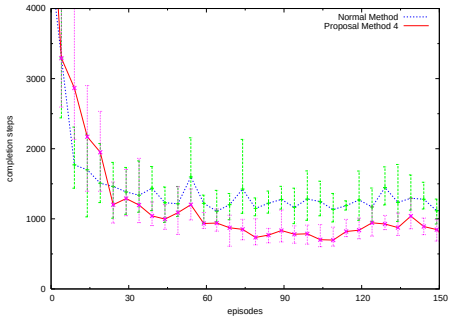

図 8: 通常の強化学習と報酬分配手法 4 の性能比較

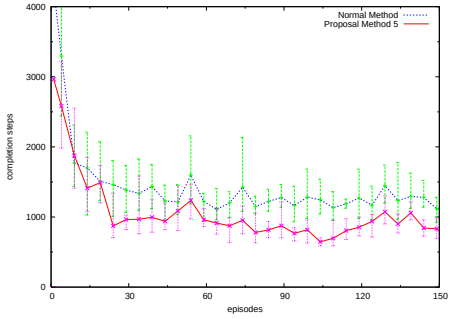

図 9: 通常の強化学習と報酬分配手法 5 の性能比較