A1-3　貢献度による報酬分配に基づくマルチエージェント強化学習

福井大学 工学部 知能システム工学科 進化ロボット研究室
池田 将之　(指導教官: 前田 陽一郎)

1　緒 言

近年、ロボットが社会で担う役割が大きくなり、多種多様な働きが求められている。それにともない、複数の自律ロボットが協力して目標を達成するマルチエージェントシステムの活躍が期待されている。マルチエージェントシステムはシングルエージェントシステムと比べ、問題解決能力、適応能力、ロバスト性などの点で利点がある [1]。しかしながら環境がより複雑になるため、学習能力をもった環境変化に柔軟に対応できるマルチエージェントシステムが必要になる。

強化学習はシングルエージェントのために開発された手法である。そのためマルチエージェントで強化学習を用いるときに、これらの強化学習をそのままを利用すると、報酬分配などの特有の問題が発生する。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 [3] ら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、各エージェントに対する貢献度を設定して、これに基づいた報酬量を決定し、貢献度に応じた報酬分配手法を提案する。またその有効性を示すため、獲物追跡問題を例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　マルチエージェント強化学習における問題点

マルチエージェント強化学習はシングルエージェント強化学習の問題点が残り、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きる。その代表的な問題について説明する。

1) 不完全知覚問題

状態空間が大きいためエージェントの知覚が限られているとき、また知覚情報が不完全のときに生じる。不完全知覚問題はシングルエージェントの課題でもあるが置かれる環境が大規模になるマルチエージェントでは免れない問題となっている。

2) 同時学習問題

エージェントが複数存在する場合、自分の行動による状態遷移先を推測するのは困難である。そして複数エージェントが独立に学習する場合、自分の学習した結果なのか、他の行動によるものか判断が困難になるため適切な学習が難しい。

3) 報酬分配問題

複数のエージェントにどのようにして報酬を分配するかという問題である。エージェントが複数の行動を行って問題を解決した場合、どの行動にどれくらいの報酬を与えるべきかという単体エージェントでの報酬分配問題と、複数のエージェントが協力して問題を解決したときどのエージェントにどれだけの報酬を与えるかという複数エージェント間での報酬分配問題がある。

3　貢献度による報酬分配法

前述の通りマルチエージェントには報酬分配問題が生じる。これはマルチエージェント系ではエージェント間の協調行動に対する報酬は定義できても、各エージェントの個々の行動に対する報酬を定義することは難しいからである。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、それがシステム全体の期待獲得報酬を上げるとは限らない。[2]、[3] の研究例は非合理的ルールではないとき全てのエージェントに報酬を与えてしまい、その報酬量も常に一定である。そのため個体の行動の評価をしているとはいえない。なぜなら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はずだからである。

そこで報酬発生するために必要な条件をエージェントの「貢献度」として求め、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。貢献度を設けることで個々の行動より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目的とする。

3.1　貢献度の求め方

マルチエージェントは複数の条件をクリアすることで報酬が発生する場合が多い。条件をクリアしたエージェントはシステム全体に対して貢献したと考えてよい。そこで条件クリアしたとき、エージェントの貢献度は時間を用いることで求める。複数のエージェントが未知の環境に置かれ、条件 $1, 2 \cdots I$ をクリアすることで報酬が発生するとする。あるエージェントが特別の感覚入力を受けたとき、つまり条件 1 をクリアしたとき貢献度 C_1 に、同様にエージェントが条件 2 をクリアしたとき貢献度は C_2 に、エージェントが行動し条件 $i(i \leq I)$ をクリアしたときの貢献度は C_i とする。条件 i をクリアした時間を T_i、条件 I がクリアされた、つまり報酬が発生した時間を T_e とする。

3.1.1　提案手法 1

提案手法 1 はシステム全体のエピソードを基準に条件達成の速さで貢献度がかわる。同じ条件をクリアする場合でも時間が早ければ早いほどその貢献度が高いと考え、貢献度は時間に比例して低くする。貢献度は C_i は式 (1) で定義する。

$$C_i = \frac{T_e - \lambda T_i}{T_e} \tag{1}$$

λ は重み係数で、範囲は $(0 < \lambda < 1)$ である。

3.1.2　提案手法 2

提案手法 2 の貢献度 C_i は T_e から始まる等比減少関数を用いる。報酬発生を引き起こしたエージェントを評価し、ステップ数が前の貢献になるほど貢献度は低くする。求める貢献度は式 (2) で定義する。

$$C_i = \lambda^{(T_e - T_i)} \tag{2}$$

提案手法１と同様に λ は重み係数で、範囲は $(0 < \lambda < 1)$ である。

3.2 貢献度に基づく報酬分配法

条件 i をクリアしたときの基本報酬量 R_i、報酬発生時に渡される報酬 r_i は式 (3) で表される。

$$r_i = R_i C_i \tag{3}$$

提案手法１の基本報酬量 R_i は条件の i の難易度やシステム全体への影響の大きさなどで決める。例えばサッカーなどの得点でシステムを評価する場合はシュートやアシストの R_i は大きくし、アシスト以前のパスやドリブルは小さくなる。また、追跡問題などの時間で評価される場合はより早く条件クリアした R_i が大きくなると考えられる。一方、提案手法２の基本報酬量 R_i は全て同じである。マルチエージェントシステム全体の大きさで決める。

ここでは学習システムは経験強化型である Profit Sharing を用いる。ここでの強化関数は等比減少関数を用いた。本来の Profit Sharing では初期状態または報酬を得た直後から報酬までのエピソード単位で学習を行う。提案手法では条件クリア時を１つの報酬のエピソードとして学習するためエージェントごとにエピソードが変わる。それにより条件 i をクリアした場合の強化関数 $f(h_i)$ は式 (4) で定義できる。

$$f(h_i) = \gamma^{h_i} r_i \tag{4}$$

報酬が得られた場合、報酬を与えられた時点を 0 ステップとし、h_i は条件 i を h ステップ前にクリアしたことを表す。また、γ は割引率である。式 (4) を用いて、状態 s の時、行動 a の評価値 $w(s,a)$ は以下のように更新される。

$$w(s,a) \leftarrow w(s,a) + \alpha f(h_i) \tag{5}$$

ここで α は学習率を表す。学習箇所を従来手法と変更することにより、無効ルールを押さえることができ、かつ学習すべき箇所を重点的に強化することができる。また条件クリアの時間を利用することでシステムの一員としての個体の行動評価をすることができ、それにともない報酬量が変化する。

4 シミュレーション実験

今回は獲物追跡問題を例題に実験を行った。本研究で用いたシミュレータは Linux 上で OpenGL のグラフィックス・ライブラリ及びその補助ライブラリである GLUT を用い、C 言語で製作したものである。

シミュレータ画面を図１に示す。ハンターエージェント (四角青色) は４体おり、それぞれ４隅に、獲物エージェント (三角赤色) は１体でフィールド中央に配置される。全ハンターエージェントの視覚、行動能力は等しい。また、ハンターエージェント、獲物エージェントは共に上下左右１マスずつ動くことができる。エージェント同士が重なる場合は動き直す。獲物エージェントは斜め方向は死角である。獲物エージェントの視覚にハンターエージェントがいる場合、最も近いエージェントから逃げる動きを選択する。２エージェントが獲物エージェントに隣接したとき捕獲したとしエピソードが終わる。条件１はハンターエージェントが一体だけ獲物に隣接する、条件２はハンターエージェントが二体獲物に隣接する設定にした。フィールドの広さ、エージェントの視覚の広さ、基本報酬量を変えながら数パターン実験を行った。

5 実験結果

フィールドの広さを 13×13、ハンターエージェントの視覚の広さは自分を中心に 5×5 とした時の環境で

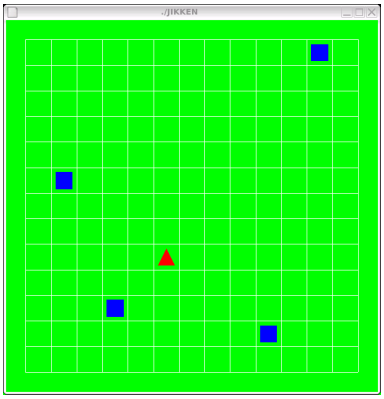

図 1. シミュレータ画面

の実験結果を示す。図 2 は 200 エピソードを 1 試行として、5 試行を行い平均をとった結果である。提案手法１は $R_1 = 100$、$R_2 = 90$、提案手法２は $R = 10$ に設定し、両手法とも $\alpha = 0.8$、$\gamma = 0.8$、$\lambda = 0.8$ で実験を行った。また、従来手法との比較のため、文献 [3] の手法も同様の実験を行った。提案手法１、２ともに獲物追跡システムとして学習できていることがわかる。両者を比較すると提案手法１が学習初期から行動回数が少なく、これは貢献度と基本報酬量の設定がうまく行われているからであると考えられる。また、初期状態では従来手法よりも提案手法１の方が行動回数が少なく、効率よく学習できていると考えられる。提案手法２については初期状態では従来手法に劣るが学習するにつれてその差は少なくなっている。

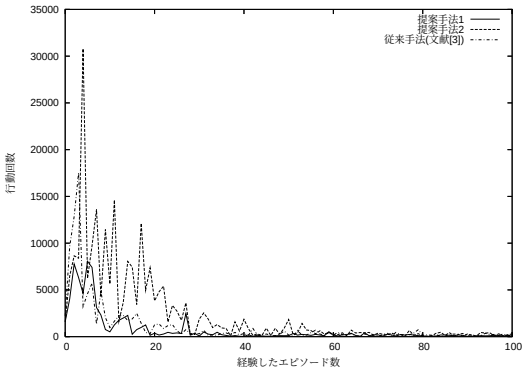

図 2. 実験結果

6 結 言

本研究では、マルチエージェント強化学習時に起きる報酬分配問題に対応できる貢献度に基づいたマルチエージェント強化学習を提案し、獲物追跡問題によりその有効性を示した。今後の課題としてエージェント数や条件を増やすなど、より複雑な環境での応用や学習速度の向上が考えられる。

参考文献

[1] 高玉 圭樹 “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配分の理論的考察,” 人工知能学会誌, Vol.13, 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.16, No.6, pp.2D3.02.1-2D3.02.4 (2002)