1L5-2

exploration 率の共有範囲によるマルチエージェント強化学習の考察

Consideration of the multi agent reinforcement learning by the joint ownership range of the exploration rate

岡野 拓哉 *1*2　野田 五十樹 *1*2*3
Takuya Okano　Itsuki Noda

*1東京工業大学 Tokyo Institute of Technology　*2産業技術総合研究所 AIST　*3JST, CREST

We investigate effects of sharing exploration ratios among agents under multi-agent reinforcement learning. In order to get optimal or better learning parameters in evolutionary ways, we need to consider the case of heterogeneous agents where each agent use different learning parameters instead of uniformed one. We conducted several experiments to measure the effects of sharing exploration ratios among agents, and measure its effects to the average learning performance. We confirmed that the average learning performance improved when sharing some degree exploration rates.

1. はじめに

マルチエージェント学習は複数の知的なエージェントの同時学習であるため、人間社会で生じる問題を様々な形で含んでいる。このことから、我々は状況により振る舞いが変化する人間により構成される社会モデルとしてマルチエージェント学習の系を取り上げる。多様なエージェント群が同時に学習を行うマルチエージェント学習では、各エージェントの学習過程がどのように相互作用するかを知ることが重要な問題である。また、マルチエージェント学習にまつわる性質を解明すると、人間社会の問題である、混雑問題や環境問題を解消する道筋につながる可能性がある。

特に重要なのが、多様なエージェントにより構成される社会の系全体の挙動である。マルチエージェント学習でいえば、学習方式に多様性がある場合に相当する。特に重要なのが、多様なエージェントにより構成される社会の系全体の挙動である。マルチエージェント学習でいえば、学習方式に多様性がある場合に相当する。従来のマルチエージェント学習の研究では、すべてのエージェントが同じ学習則及び学習パラメータを持つことを仮定していた。一方で多様なエージェントにおけるマルチエージェント学習問題は、十分に研究されていない。しかし、実世界におけるマルチエージェント学習問題では、各エージェントは様々な学習の特性を持っているため、多様なエージェントによるマルチエージェント学習問題は重要な問題である。

本研究では、exploration 率が多様な環境下であるエージェントがほかのエージェントの行動指針をまねることができる社会を取り上げ、そのような系がどのように発展していくかを、実験によって分析した。

2. マルチエージェント環境下での exploration 率

強化学習での exploration 率は学習の性能を左右する基本的な学習パラメータである。環境に一人のエージェントしかいない場合には、exploration 率は自らの行動の性質を決定するパラメータでしかない。一方、マルチエージェント環境下においては、あるエージェントは他のエージェントにとって環境の一部であるため、各エージェントの行動の性質を決める exploration 率は、すべてのエージェントに影響を与えるパラメータになる。そのためマルチエージェント環境下での exploration 率はシングルエージェントの際の exploration 率以上に様々な性質を持つ重要なパラメータである。

連絡先: 岡野拓哉, 東京工業大学, 〒 152-8550 東京都目黒区大岡山 2 丁目 12-1, 03-3726-1111, okano.t.ac@m.titech.ac.jp

2.1 exploration 率の共有

本研究では、エージェント間の情報共有、あるいはエージェントによる他エージェントの模倣を、学習パラメータの共有・コピーと見做す。エージェント間の情報共有としては、経験や学習結果の共有がまず考えられるが、その結果、前エージェントが同じ行動をとることになり、以下での取り上げる資源共有問題などでは有効に働かない。一方、学習パラメータの共有はそのような問題を生じない。現実社会のアナロジーで言えば、行動そのものを真似るのではなく、行動学習の方策を真似ることに相当する。本研究では、共有できるエージェントの範囲を拡大させていった際の系の変化について実験を行い、考察をした。具体的には、タイムステップ毎に総獲得報酬の下位グループに属するエージェント群がある一定の確率で最も総獲得報酬の高いエージェントの exploration 率をコピーすることができる環境であることを仮定する。

3. 問題設定

本節では、本研究で扱うマルチエージェント学習のゲームの一つである資源共有問題について説明する。

3.1 資源共有問題

資源共有問題とは複数の資源を複数のエージェントで共有するゲームである。タイムステップ毎に各エージェントは一つ資源を選択し、選択した資源に応じて報酬を得る。それを繰り返し行うマルチエージェントゲームである。

資源共有問題を下記のように定義する。共有する資源の集合を $R = \{r_1, r_2, \ldots, r_n\}$、資源のキャパシティを $C = \{C_{r_1}, C_{r_2}, \ldots, C_{r_n}\}$ エージェントの集合を $A = \{a_1, a_2, \ldots, a_n\}$ と定義する。

資源共有問題のゲームの流れを以下に示す。

1. それぞれのエージェント $a_i \in A$ が資源 $r_j \in R$ をそれぞれの方策に従って選択する。 ここでのエージェントの方

策は自らの利益のみを追求していく方策である。

2. それぞれの資源 r_j を選んだエージェント a_i は r_j を選んだエージェント数によってを値が変化する報酬関数 U_j に従って報酬を得る。本研究では報酬関数 U_j を以下のように定式化する

$$U_j = \frac{1}{1 + totalAgent(r_j)/C_{r_j}} \quad (1)$$

$totalAgent(r_j)$ は資源 r_j を選択したエージェント数である。

3. それぞれのエージェントは得られた報酬を元に自らの方策を更新していく。

ここでは各エージェントは強化学習によって学習しているので、エージェント a_i のは資源 r_j の期待報酬 $V_i(r_j)$ は以下の更新式により更新していく

$$V_i(r_j) = (1 - \alpha)V_i(r_j) + \alpha U_j \quad (2)$$

(i)、(ii)、(iii) を順に繰り返し行い、最終的なエージェントの平均報酬により評価を行う。

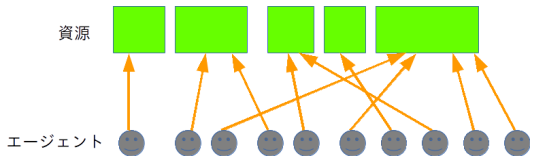

図 1: 資源共有問題イメージ図

3.2 動的資源共有問題

本研究では資源共有問題を実社会の問題に近づけるために、ある一定の確率で資源のキャパシティが変動する「動的資源共有問題」を実験に用いる。

実社会の資源共有問題の多くが動的資源共有問題といえる。実社会の動的資源共有問題を渋滞問題で例えると、資源を道路、エージェントを車としてとらえることができる。そうした場合、資源である道路は道路工事などによっていきなり通れなくなる、道幅が狭くなる、広くなるなどの資源のキャパシティが動的に変化することが考えられる。そのため、本研究では資源のキャパシティをある一定の確率により変化させる動的資源共有問題により実験を行う。

4. 実験と考察

4.1 実験設定

- ゲームの反復回数:10000 回
- エージェント数:200 体
- エージェントは ϵ-greedy により行動選択を行う。ϵ-greedy 行動選択とは ϵ の確率でランダムで次の行動を選択し環境から情報を得る (exploration)。1-ϵ の確率で今までの経験から最も多くの報酬が獲得できそうな行動をする (exploitation)
- 資源の初期設定

資源 id	0	1	2	3	4	5	6
キャパシティ	5	10	10	15	15	20	35

- 資源のキャパシティの変動：各資源はすべてのエージェントが資源を選択し報酬を受け取り、ゲームの一試行終了するごとに 0.1%の確率で資源キャパシティが変動する。具体的には、資源が変動する際にキャパシティが初期設定であれば二倍に増やす。すでに初期設定から二倍になっている資源が変動する場合には初期設定のキャパシティに戻るような変動をする。
- exploration 率の共有: タイムステップ毎に総獲得報酬が下位 x%に属するエージェントは 10%の確率で最も総獲得報酬が高いエージェントの exploration 率をコピーすることとする。この下位グループ x の範囲を拡大させていった際の全体報酬の変化について観測
- 初期の exploration 率の分布: 初期のエージェント群の exploration 率は一様分布に従った乱数により決定する。
- 理想の exploration 率: 本実験では「理想の exploration 率」を全エージェントの exploration 率が画一の時に、最大の平均報酬をとる exploration 率とする。図 3 が全エージェントが画一の exploration 率である時のゲーム終了後の全体の平均報酬の推移である。x 軸はすべてのエージェントの exploration 率である。この時に最大の平均報酬を得た exploration 率は 0.075 であった。よって、本実験では理想の exploration 率を 0.075 としている。

4.2 実験結果と考察

ここでは ϵ の分布を $[0, 1]$ の区間の一様分布に従う乱数により決定することとする。

4.2.1 exploration 率を共有できる下位グループの範囲を拡大させていった際の全体報酬の推移

exploration 率を共有できる下位グループの範囲を拡大させていった際の全体報酬の推移について調べた。この結果が図 2 のグラフである。このグラフの x 軸は exploration 率を共有できる範囲を示しており、右に行くほど exploration 率を共有できる下位グループの範囲が拡大していることを示す。また、y 軸は、各々の共有範囲設定における全体報酬の平均 (5 回の試行の平均) である。x 軸が大きくなるほど exploration 率を共有できる下位グループの範囲を拡大させていったときの全体の平均報酬の推移である。このグラフから、あるところまで共有範囲を拡大させていった際には徐々に全体の平均報酬が向上しているのがわかる。つまり、ある程度のパラメータ情報の共有は全体の学習の向上につながると考えられる。一方、共有があまりに広がりすぎると、全体報酬は下がってしまっている。これは、あまり共有しすぎるとすべてのエージェントが学習のはじめのほうでは有利な exploration 率が高いエージェントを真似てしまう。その結果より exploration 率が高いエージェントが多くなり、全体としての報酬も小さくなると考えられる。

4.2.2 理想の exploration 率と最終的に最も多くエージェントが保持している exploration 率

本節では exploration 率を共有した際のゲーム終了後の各エージェントの exploration 率の分布について分析した。2 つのことがわかった。

1 つ目は各エージェントの exploration 率は共有範囲が適切な範囲であれば、このゲーム設定では有利な低い exploration

率に集まることである。図 4 の赤いグラフが各共有範囲において最終的にどの exploration 率を持ったエージェントが最も多いかを示しているグラフである。このグラフから共有範囲が小さい時には、低い exploration 率を持つエージェントが多くなっていることがわかる。これは、低い exploration 率を持つエージェントがゲームを重ねるにつれて得をしているからあると考察できる。また、共有範囲が大きすぎるときには多くのエージェントが序盤で有利な高い exploration 率を持つエージェントを模倣する。そのため、多くのエージェントがゲームの早い段階で高い exploration 率になってしまっていると考えられる。

２つ目としては、最終的な exploration 率の分布は理想の exploration 率に一致しないことである。図 4 の緑のグラフが全体としての理想の exploration 率である。この図から exploration 率を共有させて学習行動させた時に、自律的に全体として理想である exploration 率に近づいていないことがわかる。よって本研究で用いた共有の仕方では、自律的に理想の exploration 率に収束していないといえる。

5. 終わりに

本研究では、exploration 率を共有する範囲を拡大させていった時のマルチエージェント強化学習の考察を行った。その結果、ある程度 exploration 率を共有することはマルチエージェント強化学習の性能の向上につながることがわかった。そして、あまり共有しすぎると間違った ε に収束してしまう確率が高まるため、マルチエージェント強化学習の性能の向上につながらないということを確認した。

また、エージェント間で exploration 率を共有させた際に最終的に全体としての理想の exploration 率を持つエージェントが最も多いような環境に自律的になるのか調べた。その結果、本研究で用いた共有方法では自律的に全体としての理想の exploration 率が最も多くのエージェントが保持しているような環境にはならないことを確認した。

これらの現象は、パラメータの多様性維持ができなかったことに起因すると考えられる。今回用いた実験設定では、パラメータの共有を最も成績の良いエージェントに限ってしまっていた。このため、たまたま初期に良い成績を収めたエージェントのパラメータが流布してしまい、学習パラメータの探索として適切な exploration が行えない。このため、今後はこの多様性の維持を含めた実験・分析を進めていく必要がある。

参考文献

[Noda 13] 野田 五十樹:動的環境におけるマルチエージェント同時学習に関する考察 (2013).

[Okano 15] 岡野 拓哉、野田 五十樹:多様な exploration 率を持ったマルチエージェント強化学習の考察 (2015).

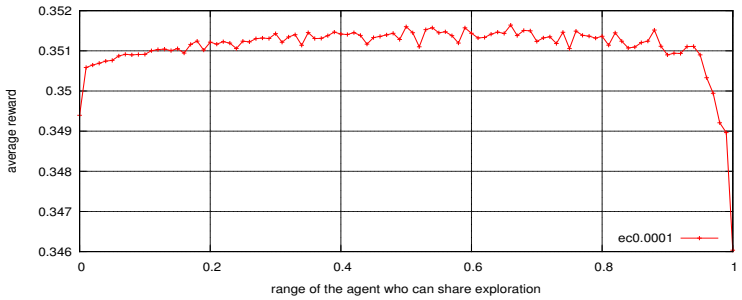

図 2: explaration 率を共有できるエージェント群の範囲を拡大させていった際の試行後の平均報酬の推移。x 軸は下位グループの範囲を示している。0.1 であれば下位 10%のエージェント群がタイムステップ毎に 10%の確率で最も得をしているエージェントの ϵ を共有する

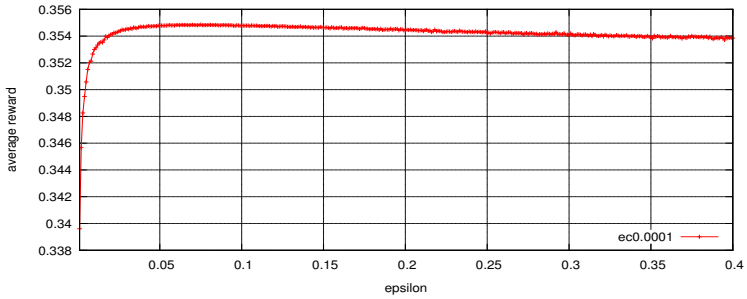

図 3: 全エージェントの exploration 率が画一の時の平均報酬

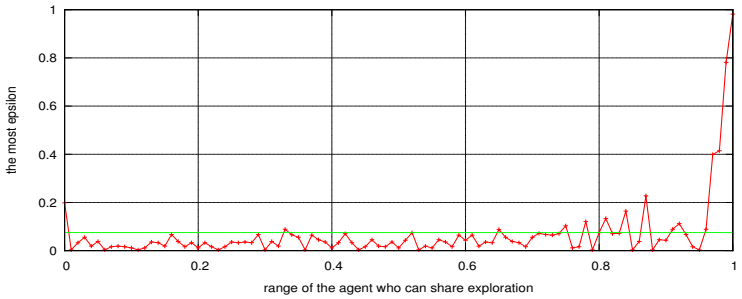

図 4: 理想の exploration 率と各共有範囲でゲームを行った際の最終的に一番多くのエージェントが保持していると exploration 率の関係を表したグラフ。緑のグラフが理想の exploration 率であり、赤のグラフが各共有範囲で最終的に最も多くのエージェントが保持していた exploration 率である。