F-035

マルチエージェント強化学習における行動選択手法
A Method of Action Selection in Multi-Agent Reinforcement Learning

渡辺　潤† 延澤志保‡ 太原育夫‡
Jun Watanabe Shiho Nobesawa Ikuo Tahara

1. はじめに

強化学習は相互作用に基づく目標指向型の学習であることが特徴的である．エージェントは試行錯誤を通じて最適な政策を見つけることを目標とする．強化学習では完全な環境モデルが必要でないことから，マルチエージェントの学習の枠組みとして最適であると考えられている．しかし，マルチエージェントに強化学習を適用する場合，状態空間が膨大であるため，学習に時間がかかるという問題点が指摘されている．本稿ではマルチエージェントの環境で，εグリーディ手法とルーレット選択の振る舞いの違いについて考察し，それを用いて新たな行動選択手法を提案し，学習の速度の向上に有効であることを示す．

2. 強化学習

2.1 Profit Sharing

強化学習手法として代表的なものに Profit Sharing がある．Profit Sharing は学習の速度が速く，非マルコフ的な環境でも頑健であり，合理性定理より，無効ルールを抑制することができるため，マルチエージェント強化学習の手法に最適と言われている [1]．エージェントはそれまで認識した状態と行動の組（ルール）を記憶し，報酬をそれぞれのルールの重みとして分配する．i ステップ目のルールの重みを w_i，報酬を r，エピソード長を W，エージェントの行動数を $|A|$ で表すと，合理性定理を適用した Profit Sharing は

$$w_i \leftarrow w_i + f_i\ ,\ f_i \leftarrow \frac{1}{|A|} f_{i+1}\ ,\ f_W = r$$

と表すことができる．

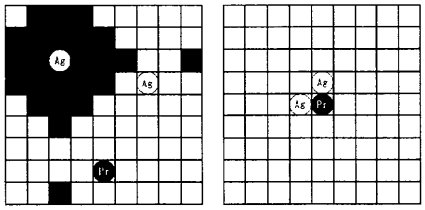
2.2 行動選択手法

行動選択手法とは，学習の結果得られたルールの重みから，適切な行動を選択する手法である．たとえ良い学習で良いルールの重みを手に入れても，適切な行動選択手法をとらなければ，エージェントは合理的な行動をとることができない．環境に応じ，適した手法を採用することで，学習の効率に影響を及ぼす．

3. 行動選択手法による学習の違い

3.1 実験内容

εグリーディ手法とルーレット選択のマルチエージェント強化学習での振る舞いの違いを調べるために次の実験を行った．エージェント 2 体，獲物 1 体を 9 マス×9 マスのフィールドにランダムに配置する（図 1）．フィールドには壁の概念がなく，フィールドの端はループしている．エージェントと獲物は交互に 1 マスずつ移動する．エージェントは 2 体同時に獲物の 1 マス隣に移動できれば，タスクが達成され報酬を受け取る．エージェントは自分を中心に，図 1(a) の網掛け部分にいる獲物とエージェントを認知できる．獲物はランダムに移動する．



(a) 初期状態　(b) 終了状態

図 1: 追跡問題

3.2 実験結果

εグリーディ手法とルーレット選択で学習させた結果を図 2 に示す．

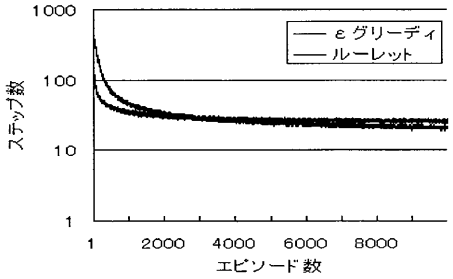


図 2: 行動選択手法による学習の違い

図 2 からわかるように，εグリーディ手法は収束までの時間は短いが，収束後の 1 エピソードの平均ステップ数はルーレット選択に比べて長い．一方ルーレット選択は収束までの時間は長いが，収束後の 1 エピソードの平均ステップ数はεグリーディ手法に比べて短いことが示された．

4. 切り替え手法を用いた強化学習

4.1 切り替え手法

強化学習では一般的に，学習前半では広く探索をし，後半では合理的政策を実行することが好ましいとされている．

学習前半では広く探索をすることが要求されているが，マルチエージェントのような，膨大な状態空間が存在する環境下では，広く探索を行うと莫大な時間がかか

†東京理科大学大学院理工学研究科情報科学専攻
‡東京理科大学理工学部情報科学科

る．特に長いエピソードで合理性定理を適用したProfit Sharingを用いた場合，エピソード初期のルールはほとんど強化されないため，長い時間探索しても，高い効果は期待できない．従って探索を効率的に行うことによって，学習の効率化が期待できる．εグリーディ手法の場合，現段階までに学習された政策の中で，一番効率のよい政策を高確率で選択することが可能である．その一番効率のよい政策の周辺を探索するので，探索効率が期待できる．従って学習前半ではεグリーディ手法が有効であるといえる．

学習後半では合理的に行動することが要求されているが，曖昧性が強いマルチエージェント環境下では，合理的な行動が1つに決定されるとは限らない．図3を例としてあげる．黒丸をエージェントのいる位置とする．点線内でエージェントは不完全知覚を起こし，点線内の状態は全て同一の状態と認識する．(a)の場合，遷移先でどちらの行動を選択しても報酬にたどり着けるが，(b)の場合，次の遷移先によって合理的な行動が変わってくる．エージェントはどちらが合理的な行動かがわからないため，どちらも行動できる余地を残さなければならない．このことよりグリーディな行動選択は危険性がある．従って，学習後半ではルーレット選択が有効であるといえる．以上より，ある程度学習が収束するまでεグリーディ手法を用い，収束後ルーレット選択を用いれば，よりよい行動選択手法になることが期待できる．本稿では学習途中で行動選択手法を切り替える切り替え手法を提案する．

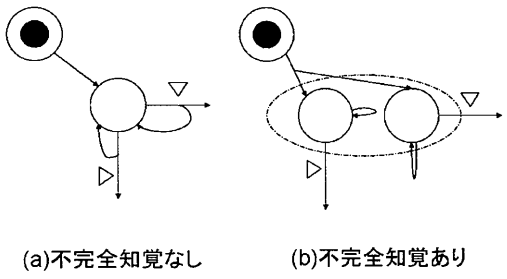


図3: 行動選択における曖昧性

4.2 学習残エントロピー

学習を行う前に切り替える最善のタイミングを取得することは困難であるので，学習途中でそのタイミングを取得することが必要である．そのためには学習の進行度を表す指標が必要である．この指標として学習残エントロピー[2]を用いる．$|A|$を行動の種類，$|E|$を状態の総数，$p(s,a)$を状態sのとき行動aを選択する確率とすると，学習残エントロピーは

$$I(s) = \frac{1}{|A|}\sum_{a\in A(s)} p(s,a)\log(p(s,a))$$
$$I = \frac{1}{|E|}\sum_{s\in S} I(s)$$

と定義される．学習残エントロピーは本来，行動の曖昧さを測る指標である．学習初期では行動は一意に定まらず学習残エントロピーは高いが，学習が進むにつれて行動の曖昧さは解消され学習残エントロピーは低くなることより，学習残エントロピーは学習の進行度を測る指標として用いることができる．学習残エントロピーの計算は全状態をスキャンするため，計算量は多いように見えるが，更新したルールのみを考慮すればよいので，計算量は少なく抑えられる．

4.3 実験内容と実験結果

切り替え手法は，学習残エントロピーの値を観測し，ある閾値で行動選択手法を切り替える手法である．切り替え手法を用いた実験として，3.1節と同様の実験を行った．切り替え手法の閾値は0.5と定めた．図4に実験結果を示す．

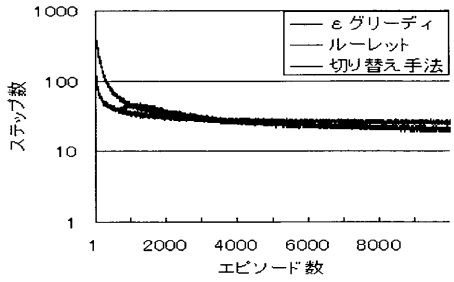


図4: 切り替え手法における実験結果

図4より切り替え手法は総ステップ数だけでなく，収束後のステップ数も一番良い結果を示していることがわかる．収束後のステップ数はεグリーディ手法では27ステップ，ルーレット選択では21ステップ，切り替え手法では19ステップかかった．この理由として，切り替え手法では，前半にεグリーディ手法を行うことによって，ルールの重みに多少の偏りを起こすことができる．このことによって，ルーレット選択にて合理的な行動を選択させやすくなるため，ステップ数を下げることができたと考えられる．パラメータを変えても同様の結果が得られた．

5. まとめ

本稿ではマルチエージェント強化学習においての行動選択手法による学習の違いを示し，それらを用いてマルチエージェント強化学習に効果的な行動選択手法として，切り替え手法を提案し，その有効性を示した．

今後の課題として，より環境に強固な学習の進行度の指標の導入や，パラメータを用いたシームレスな切り替え手法の提案などが挙げられる．

参考文献

[1] 宮崎和光，木村元，小林重信，“Profit Sharingに基づく強化学習の理論と応用，”人工知能学会誌，Vol.14, No.5, pp.800-807 (1999)

[2] 伊藤昭，金渕満，“知覚情報の粗視化によるマルチエージェント強化学習の高速化 —ハンターゲームを例に—，”電子通信学会論文誌，Vol.J84-D-I，No.3, pp.285-293 (2001)