F-012

粗視化を用いないProfit Sharingによる強化学習の効率化

Efficient Reinforcement Learning by Profit Sharing without Coarse Graining

細井 健輔† 松井 丈弥† 能登 正人†
Kensuke HOSOI Takeya MATSUI Masato NOTO

1. はじめに

近年，複数の自律的なエージェントの行動により複雑かつ動的な問題を解決しようとするマルチエージェントの研究が盛んに進められている [1]．マルチエージェントシステムを構成するエージェントの設計方法のひとつに強化学習がある．強化学習の代表的な手法として，Profit Sharing がある．Profit Sharing は状態数が大きい場合は学習が困難であり，学習の効率が著しく低下する．この問題を解決するために，観測情報の一部を同じ状態であるとみなすことにより状態数の削減を図る「粗視化 (Coarse-Graining)」がよく用いられている．しかし状態数の削減を大幅に行うことは，試行錯誤の意義やエージェント間の協調性を失いかねない．また，問題ごとに最適な粗視化を設計することは，設計者にとって大きな負担になる．そのため，粗視化に頼らない手法を提案する必要がある．

本研究では，粗視化を用いず蓄積された報酬を有効に活用し，割引率に工夫を加えることで学習を高速化，効率化する手法を提案する．提案手法では割引率が小さい場合に起こる学習の遅延と，割引率が大きい場合に起こる経験固執の問題の改善が期待できる．従来手法と提案手法を追跡問題に実装し，その結果の比較を行う．

2. Profit Sharing

Profit Sharing は正の報酬 r_t を獲得した時，エピソード内の各ルールに報酬 r_t の一部を分配し，累積することで強化を行う．各ルールに累積された報酬値を評価値と呼び，ルール選択時の判断基準に使われる．評価値の更新は式 (1) のような等比減少関数による更新式を用いる．

$$V(s_t, a_t) \leftarrow V(s_t, a_t) + \gamma * f(x) \tag{1}$$

ここで t は現在時刻を表し，$s_t, a_t, V(s_t, a_t)$ は時刻 t での状態，行動，評価値で，γ は基本報酬である．$V(s_t, a_t)$ の値が 0 であると，経験固執問題の典型的なモデルとなってしまうため，初期値として α を与える．強化関数 $f(x)$ は一般に等比減少関数 $f(x) = (1/L)^x$ とされ，L の値は各状態におけるルールの数で十分であることが証明されている．目標からさかのぼって分配するため，関数 $f(x)$ の引数 x は時刻 t と逆方向になる．初期値，基本報酬は任意のパラメータで与えられ，適応する問題ごとに最適な値は異なる．この更新式より，時刻 t が小さくなるほど報酬が割り引かれ，報酬が与えられる直前の行動ほど高い報酬が与えられる．このような更新式で報酬を与えるのは，エピソード終了直前の行動ほど問題解決に貢献した強化すべき行動という考えからである．

状態 s における行動 a の決定は，状態 s の評価値 $V(s, a)$ を要素としたルール選択法を用いることが多い．一般的に，評価値の比率に応じてルーレット選択を用いて行われる．これにより，評価値の高い行動ほど採択される確率が高くなり，低い行動ほど採択されにくくなる．ルーレット選択により，ある状態 s において次ステップで行動 a が選択される確率 p は次の式 (2) で表される．

$$p(s, a) = \frac{V(s, a)}{\sum^{a'} V(s, a)} \tag{2}$$

3. 提案手法

充分に学習の進んだ状態で選択する行動はほぼ決まっており，高い確率で「良い選択」をする．良い選択をすることは，目標状態に到達するためには不可欠である．つまり「良い状態への選択は良い選択」と言え，良い選択をした場合には，大きい報酬を与えるべきである．このことから報酬値の分配の際に「もし，その状態の報酬値が，初期値の h 倍であれば，報酬値の割引を行わない.」という条件を加え，学習を行う（以下，提案手法 1）．しかし，この条件が原因となり，報酬が過剰に大きくなる可能性があるため，さらに「その行動の選択肢の確率が一番高いか？」という条件を加え，学習を行う（以下，提案手法 2）．これらの手法を用いることにより，学習が進むほど，目標状態から遠い状態にも，大きい報酬を分配することができる．そのため，状態数が大きい場合でも効率的な学習ができることが見込まれる．

4. 実験方法

4.1 追跡問題

追跡問題とは，複数のハンターが 1 人の逃走者を追いかけ，捕獲することを目標とする標準問題である．

本研究ではハンターをエージェントとし，$N \times N$ の空間にハンターを 2 体，獲物を 1 体配置する．この状態からハンターと獲物は上下左右のいずれかに移動し，獲物はランダムに移動する（図 1）．ハンターが獲物を上

†神奈川大学大学院工学研究科電気電子情報工学専攻，Graduate School of Electrical, Electronics and Information Engineering, Kanagawa University

下，または左右ではさんだ状態（図 2）を捕獲と定義する．また状態数が大きくなる値として $N = 15$，初期値 α を 10，基本報酬 γ を 100 とした．

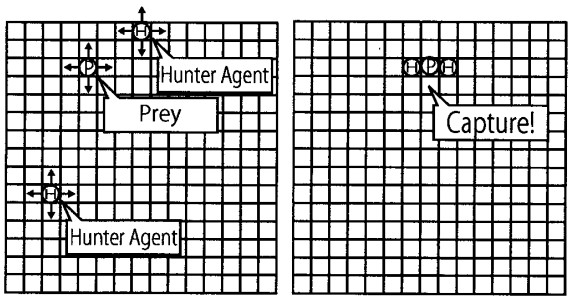

図 1: 追跡問題:追跡中

図 2: 追跡問題:捕獲

4.2 粗視化

粗視化には，観測情報全体を簡略化する方法 [2] と，目的に必要な情報のみを採用する方法 [3] の二種類が提案されている．追跡問題で扱う粗視化は，位置情報を上下左右とその中間をあわせた 8 方向に分類し状態数を削減する場合と，お互いのエージェントの位置情報を無視する場合が考えられる．お互いのエージェントの位置情報を無視する場合は，もはやマルチエージェント環境下の問題とはいえず，単に問題の難度を計るだけになってしまうため，本研究では位置情報を制限する場合を考える．

本研究で扱う粗視化として図 3 と図 4 を用意した．図 3 は，追跡問題には非常に効果的である粗視化（以下，粗視化 1）である．図 4 は，距離が離れた部分は知覚が制限されるという現実的な粗視化（以下，粗視化 2）である．この両粗視化と従来手法，提案手法を比較する．

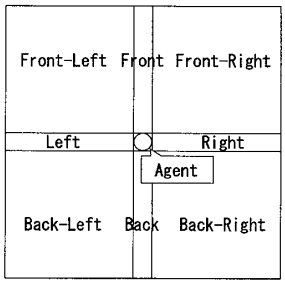

図 3: 粗視化 1

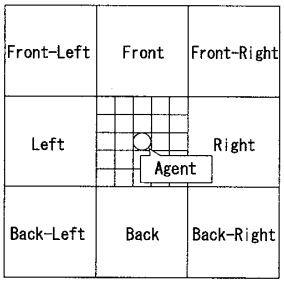

図 4: 粗視化 2

5. 結果と考察

提案手法 1，提案手法 2，従来手法，粗視化 1，粗視化 2 を適用した場合の追跡問題のシミュレーション実験の結果を図 5 に示す．

学習速度については粗視化 1 が早く，追跡問題において効果的な設計であることがわかる．従来手法と粗視化 2 は収束値に差はあるが，ほぼ同じ速度で学習が進んでいる．これは，どちらもいかに目標状態から遠い状態に報酬値を与えるかを考えて設計されたからである．また提案手法 1，2 はほぼ同じ速度で学習が進んでいる．これは，h の値を学習に充分な値を適用し，ある程度従来手法で学習を進めた後，提案手法を適用することで一気に学習を進めたからである．

各手法の最終的な捕獲ステップ数は 提案手法 2 < 粗視化 1 ≤ 提案手法 1 < 粗視化 2 < 従来手法 となっている．この結果から，提案手法 2 は粗視化 1 よりも良い結果となり，提案手法 1 においても粗視化 1 と僅差である．提案手法は人間が設計した粗視化よりも良い結果，もしくは同等の結果出せることがわかった．

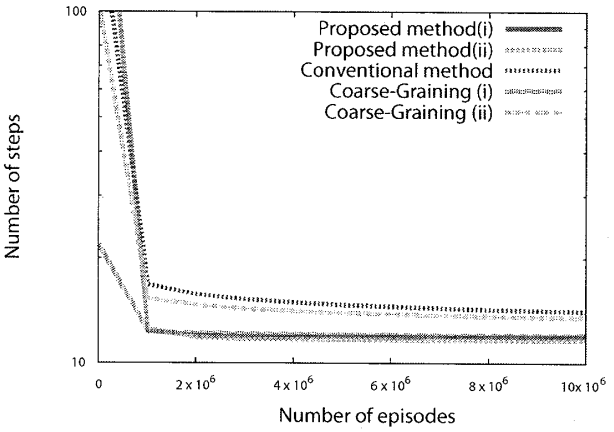

図 5: 学習速度の比較

6. おわりに

本研究では，割引率に工夫を加えることで，状態数が大きい場合でも，学習を進められることがわかった．しかし，現状では良い状態の判断に初期値の h 倍という判断基準を設けているが，これも自由度の高いパラメータなので，設計時の負担を増やしてしまう．この問題を解決することが今後の課題である．

参考文献

[1] 植村　渉，上野敦志，辰巳昭治：経験に固執しない Profit Sharing 法，人工知能学会論文誌， Vol. 21, No. 1, pp. 81–93 (2006).

[2] Sutton, R.: Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding, *Advances in Neural Information Processing Systems*, Vol. 8, pp. 1038–1044 (1996).

[3] Ono, N. and Fukumoto, K.: Multi-agent Reinforcement Learning: A Modular Approach, *Proc. of the 2nd International Conference on Multiagent Systems, AAAI Press*, pp. 252–258 (1996).