LF-007

交渉過程を導入したマルチエージェントシステムにおける報酬分配学習
Reward Distribution Learning During Bargaining in a Multi-agent System

鶴岡　久†
Hisashi Tsuruoka

1. まえがき

マルチエージェント強化学習において、各エピソードの終わりで環境から与えられる報酬をエージェントに割り振る問題は報酬分配問題と言われている。従来、profit sharing のような経験強化型強化学習にはシングルエージェントを対象に報酬に貢献しない無効ルールの抑制や報酬からどれだけ過去かを引数として強化値を返す強化関数に関する研究[1]があるが、マルチエージェントを対象にした報酬分配や強化関数の研究例は少ない[2]。

環境同定型の強化学習ではマルチエージェントを対象に学習の収束を見ながら、分配率を調節し、最適値を探索する型のアルゴリズムが提案されている[3]が、探索時間が課題と考えられる。

ここでは環境同定型強化学習アルゴリズムを対象に、報酬が発生した時点でエージェン同士が交渉によって報酬分配し、交渉過程を強化学習アルゴリズムに組み入れることにより、報酬分配行動を同時に学習させ、分配率の自動探索とエージェントの学習時間の短縮を目指す。

2. 交渉過程を導入した報酬分配

本提案では報酬が発生した時点でエージェントが交渉過程に入る。ここで定義する交渉は互いに報酬分配に対する態度（例えば独占、半独占、協調（遠慮）の3タイプ）を表明することである。提示された態度の組み合わせから、あらかじめ定めた利得表に従って報酬分配する。互いの報酬の期待値を最大化するよう、その都度適切な態度を表明しなければならないが、このようなプロセスも行動の学習と同様強化学習で扱う。報酬分配機構は各学習エージェントに組みこまれる。

以下2ハンターと学習しない1獲物の捕獲問題で提案方法を説明しよう。ハンターは1匹でも獲物を捕獲できるとする。利得分配表の一例を表1に示す。表の横軸は獲物を捕らえた主ハンターの３つの態度、縦軸は捕らえ損なった従ハンターの態度を示し、数字は主ハンター、従ハンターの獲得率を示す。この表では１匹で捕獲した事実から主ハンターの取得率をやや高く設定している。態度の種類を増加させれば、分配率は細かく設定できる。

表１では（独占、独占）はナッシュ均衡でパレート最適でもある。ゲームを１回限りとすれば、捕えた獲物を独占するのが得をすることになるが、この問題はハンターの累積報酬の最大化を目指して、捕獲ステップが最小値に収束するまで繰り返えされる繰り返しゲームである。囚人のジレンマゲームも１回限りのゲームでは裏切り行為がナッシュ均衡解であるが、繰り返しゲームで協力解の出現が期待される。

†福岡工業大学、Fukuoka Institute of Technology

表１　利得分配表の一例
Table 1 Example of a pay-off table

従 ＼ 主	独占	半独占	協調
独占	(0.8,0.2)	(0.6,0.4)	(0.4,0.6)
半独占	(0.9,0.1)	(0.7,0.3)	(0.5,0.5)
協調	(1,0)	(0.8,0.2)	(0.6,0.4)

さらに相手行動や相手の報酬分配政策を予測することで、独占という戦略に勝るどのような報酬分配の動的変化が生まれるか、分配率を学習中一定にした場合に比較して最短捕獲ステップを獲得するまでの学習速度は速まるかどうか、を調べる。

交渉動作のモデルは有限オートマトンで表現する[4]。ハンターの態度の履歴を状態ととらえ、捕獲が生じた場合の態度を行動と考える。ここでは１次マルコフ過程を考える。

相手ハンターの交渉方策を固定したとき、自己の状態Ｓ、自己の態度 a^m, 相手態度 a^o としたときの交渉動作の価値関数を $QC(S, a^m, a^o)$ で表す。a^m,a^o の組で利得行列から得られる報酬を $r(a^m, a^o)$ とすれば、

$$QC(S, a^m, a^o) = (1-\alpha)\, QC(S, a^m, a^o) + r(a^m, a^o) + \gamma \max_{am} QC(S', a^m)$$

$$QC(S', a^m) = \Sigma_{ao}\, p(a^o \mid S', a^m)\, QC(S', a^m, a^o)$$

$p(a^o \mid S', a^m)$は状態 S'で自己が態度 a^m をとったときの相手態度 a^o の確率である。$p(a^o \mid S, a^m)$は状態 S での態度の生起確率を観測することで求められる。時刻 t で状態 St にあり、行動$(a^m(t), a^o(t))$が観測されたとき

$$p(a^m, a^o \mid St) \leftarrow p(a^m, a^o \mid St) + \delta \quad (a^m = a^m(t), a^o = a^o(t) \text{のとき})$$

$$p(a^m, a^o \mid St) \leftarrow (1-\delta)\, p(a^m, a^o \mid St) \quad (\text{otherwise})$$

$$p(ao \mid St, am) = p(a^m, a^o \mid St) / \Sigma ao\, p(a^m, a^o \mid St)$$

により確率が更新される。

上記の学習は捕獲動作の学習でも同様に行なわれる。

3．実験結果

追跡問題を獲物の行動方策から１）ランダム行動、２）通常はランダムに動くが、いずれかのハンターが距離 λ 以内に接近したら、ハンターと同方向へ逃げ、２匹のハンターが一定距離以内に接近した場合はハンターと９０度もしくは１８０度方向へ逃げる、の２種類を考える。後者は λ の値により、ハンター同士の協力の必要度が変わると思われる。

また両ハンターの行動能力から a) 両ハンターの行動が左右上下停止の５行動（均質ハンター）b)片方が５行動でもう一方のハンターが３行動（上下停止）（ヘテロハンター）に分類する。

上記分類を組み合わせ、報酬分配率の最適値が予想しにくい問題として、２）a)で λ＝３の場合（実験Ⅰ）、１)b)の組み合わせ（実験Ⅱ）、２）b)で λ＝１０の場合（実験Ⅲ）を取り上げ、提案方法の効果を確認する。学習

率　α=0.3、割引率　γ=0.9、フィールドは７×７（トーラス）とする。

実験Ｉの結果を図１に示す。交渉方式の２０回平均学習時間 T_{ave}(average learning time)は分配率を固定した場合の最適な分配率０．６にほぼ等しい。実験Ⅱでは図２に示すように提案方式の T_{ave} は分配率を固定した場合の最適な分配率０．９に近い。実験Ⅲの結果を図３に示す。分配率を固定した場合の最適な分配率０．６と交渉方式の学習時間 T_{ave} はほぼ等しい。

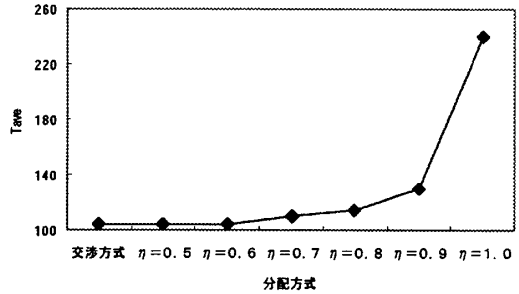

図１均質ハンター、獲物方策変化の学習時間

Fig.1Differences in T_{ave} for various distribution method (Non-stationary prey)

以上のように報酬分配率の予想が難しい場合においても提案方法では最適固定分配率にほぼ等しい学習時間で捕獲学習が行なわれることが示された。図４は図１の交渉方式の実験例で、両ハンターの交渉方策の変化を捕獲ステップＰＸの収束過程に重ねたものである。縦軸１６０、１４０、１２０が協調、半独占、独占に対応している。この例では態度の動的変動が続いているが、平均値としては最適固定分配率に収束している。

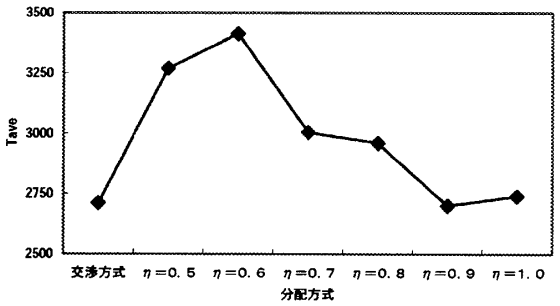

図２．ヘテロハンター、獲物ランダム行動の学習時間

Fig.2 Differences in T_{ave} for various distribution method (Hetero- hunter and random prey behavior)

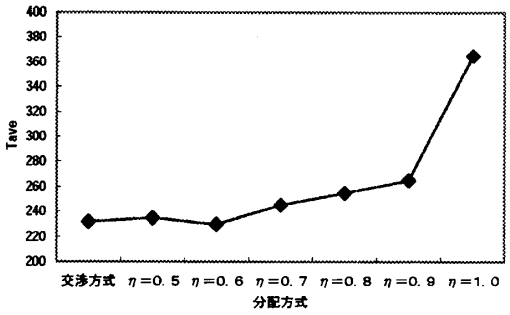

図３．ヘテロハンター、獲物方策変化の学習時間

Fig.3 Differences in T_{ave} for various distribution method (Hetero-hunter and non-stationary prey)

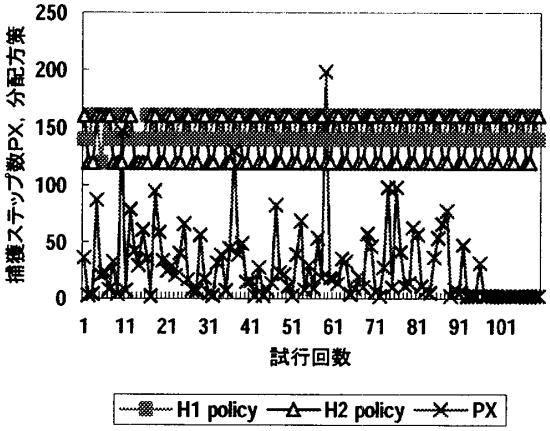

図４．ハンターの交渉態度の変化

Fig.4 Change of bargaining behavior

４．結言

本論文では、各エージェントに交渉価値関数をおき、報酬に対する態度を選択することにより、あらかじめ定めた利得分配表に従って報酬分配する。こうして報酬分配行動とエージェントの行動学習を並行して学習することにより、あらかじめ分配率を固定的に定めた最適値にほぼ等しい平均学習時間が得られ、分配率の自動最適値探索が可能であることを２ハンター、１獲物の逃走ゲームで確認した。高次マルコフ交渉過程の利用、適格度トレースへの適用等の検討が今後の課題である。

５．参考文献

１）宮崎和光、山村雅幸、小林重信: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.9,No.4,pp.104-111(1994)

２)三上貞芳：強化学習のマルチエージェント系への応用、人工知能学会誌、Vol.12, No.6pp.37-41(1997)

３）柴田克成、真崎勉：多人数ゲームにおける報酬分配学習、計測自動制御学会システム情報部門講演会、pp15-20(2002)

４）伊藤　昭、水野将史、松本達明、寺田和憲：マルチエージェント強化学習による交渉問題へのアプローチ", 信学技報,ＡＩ２００３－８４, pp19-24 (2004)