3T-5

相互作用型階層強化学習システムを用いたエージェント集団の共存期間伸長

山﨑 大地† 北越 大輔† 鈴木 雅人†
†東京工業高等専門学校

1 はじめに

マルチエージェントに関する研究分野において，エージェント間の利益格差が時間とともに拡大することで集団が共存不可能となる状況が，“格差拡大状態”として定義されている．当該状態に対して，利益格差を解消することで，集団全体の利益最大化を目指す相互作用型階層強化学習システム（Interactive Hierarchic Reinforcement Learning system：IH-RL）が提案されている[1].

これまでの研究成果から，IH-RLは共存終了に繋がるいくつかの問題点を解決し，エージェント集団の共存期間を伸長可能であることが確認されている．一方，エージェントの状態によっては利益格差が拡大してしまうような状況に陥る，といった問題も発生している[2]. 本稿ではIH-RLの設定について再検討し，利益格差のさらなる軽減・解消および，それに伴う共存期間の伸長を図る.

2 格差拡大状態

マルチエージェント環境において，各エージェントが達成すべき目標と集団全体の目的との間には，様々な関係性が想定される[3]. また，文献[1]では，各個体に利益格差が生じている状態が長期間継続することで，集団全体の利益が損なわれる“格差拡大状態”が定義されている．例えば，肉食動物の群れの中の各個体は，多くの獲物を捕えることを自身の目的として行動するが，集団全体の目的の一つは群れ全体が長期間共存することであると考えられる．獲得できる獲物の数（狩猟能力）に差が生じると，長期間獲物を獲得できない個体は死亡するため，集団全体の利益は大きく損なわれることになる．当該状態において集団の利益を最大化するには，状況に応じて各個体の振舞を，「格差」が減少する方向に変化させる必要がある．文献[1]ではエージェントの活動可能な残り時間を“寿命”，集団中の全エージェントの寿命が0となることなく活動可能な状況を“共存”と定義している．寿命はエージェントが報酬を獲得できない期間に比例して減少する.

3 相互作用型階層強化学習システム（IH-RL）

本節では，IH-RLにおける基本的枠組について述べる.

3.1 IH-RLの概要

IH-RL では複数の Playing Agent（PA）と一体の Management Agent (MA) が存在し，それぞれ Profit Sharing, Bucket Brigade と呼ばれる強化学習法によって振舞を学習する（Fig. 1）．各 PA は環境から自身の状態を観測し，自身の利益最大化を目的として，独立に適切な振舞を学習する．MA は各 PA の状態を自身の状態として観測し，PA 集団全体としての利益を最大化するような命令を 1 体の PA に対して出力する．当該 PA は与えられた命令をもとに自身の振舞を修正する．MA は自身の行動選択の結果，集団の振舞が改善

Interactive Hierarchic Reinforcement Learning System to Extend the Period of Coexistence for Agents
†Daichi Yamazaki, †Daisuke Kitakoshi, †Masato Suzuki
†Tokyo National College of Technology

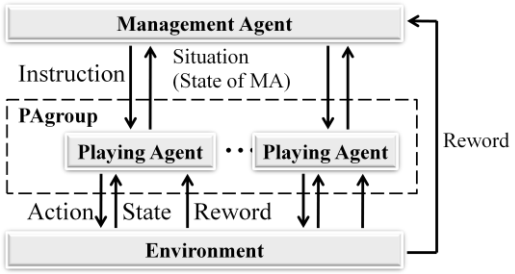

Fig.1 Framework of IH-RL

したか否かに関する情報（報酬）を次の行動選択時に与えられ，その結果をもとに適切な振舞（命令の与え方）を学習する．このようにIH-RLでは，MAとPA双方の学習結果を相互利用し，相乗効果による学習性能の向上を目指す.

3.2 従来のMAにおける学習設定

文献[2]におけるIH-RLでは，PAが報酬を獲得するか，行動を一定回数選択するまでを 1TP，PA 集団の共存が終了するか50000TPが経過するまでを1TMと定義している．MAの状態は，直近 50TP における各 PA の報酬獲得数をもとに特定され，ルーレット選択によって，50TP 毎に特定の PA_i に対して以下に示す命令を選択し，行動として出力する.

- 待機：なにもしない
- 平滑化：ルール重みを重み全体の平均値に設定
- 係数増加：報酬獲得係数ζ_iを0.05増加（$\zeta_i \in [-2,2]$）
- 係数減少：報酬獲得係数ζ_iを0.05減少

平滑化行動には，指定したPA以外のPAに報酬獲得の機会を与える効果がある．また，係数の増減行動では PA の学習速度を変化させることができる．係数増減行動は，PAのルール重み更新式を以下の式とすることで実現される.

$$\omega(s_t, a_t) = \omega(s_t, a_t) + f_t r \cdot \zeta_i \quad (t = 0, \ldots, T) \qquad (1)$$

MA の報酬獲得条件は，MA の n 回目の行動選択までにおける PA 集団の平均報酬獲得数N_nと各 PA の報酬獲得数の分散σ_n^2について，$N_n - N_{n-1} \geqq 0$かつ$\sigma^2{}_n - \sigma^2{}_{n-1} < 0$となった時と定義されており，PA 集団の利益を増加させつつ格差を解消するような行動に対して報酬が与えられる[2]．MA のルール重みの初期値および最小値は$\omega_0 (= 0.1)$とする.

4 MAの学習設定の改良

3.2節のようなMAの設定においてPA集団の共存が失敗する原因について調査したところ，以下の問題点が確認できた.

(1) MAが不適切な振舞を学習することで，利益格差が解消できずにPAの寿命が0になる
(2) 係数減少によってPAの報酬獲得係数＜ 0となることで，報酬獲得に貢献する行動が学習不能となる

これらを踏まえ本稿では，PA集団の共存期間伸長のため，MAの学習設定を改良する.

従来のMAの報酬獲得条件はPA集団の振舞の些細な変化に対しても達成可能な上，報酬未獲得時もBucket Brigadeにより様々なルールの重みが増加していた．これに伴い，適切な命令の選択率が相対的に低下することで問題(1)が発生し

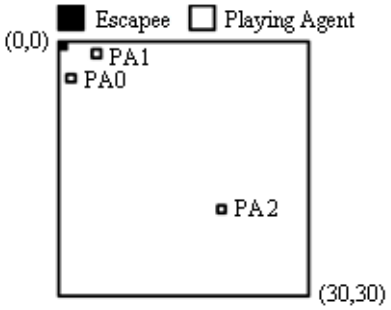

Fig.2 Simulation environment

ていた．そこで，PA 集団の利益獲得および格差軽減を並列に評価し，共存期間伸長に繋がる命令のみを MA に学習させるため，各 PA の報酬獲得数を掛け合わせた評価値 GR が閾値 T_{GR} を超えた場合に報酬を与えることとする．また，いずれかの PA の寿命が減少して集団が共存終了に近づいていると考えられる状況（GR=0）では負の報酬を与え，不要なルールの学習を抑制する．問題(2)については，報酬獲得係数の下限を 0 とすることで対応する．さらに，PA 集団が容易に適切な振舞を獲得可能な環境を想定すると，係数が正ならば学習速度の大きな変化は生じないと考えられるため，報酬獲得係数の増減行動を「ζ_i を 1 もしくは 0 に固定する」と変更する．

その他の問題点として，MA の試行後期における平滑化実行により，PA の全ルール重みが等しく高値となるため，行動の選択確率が一様になり，これ以降当該 PA の学習が困難になる問題も確認している．一般的に，共存に成功しているのであれば，各 PA の獲得利益は高い方が望ましいと考えられるため，上記の問題を解決すべく，MA の平滑化行動ではルール重みを初期化するように変更した．また，文献[2]における MA の命令間隔は 50TP となっているが，命令の頻度は可能な限り多い方が望ましい上，命令間隔を 20TP 程度まで減少させても MA の状態認識に影響はないと考えられるため，MA の命令選択間隔を 20TP に変更した．

5 計算機実験

本稿では，エージェント追跡問題を対象として提案システムの性能評価を行う．当該問題では，追跡者が逃避者を捕獲することを目的とする．追跡者を IH-RL における PA，PA 集団に命令を与える上位エージェントを MA とする．実験環境および各エージェントの初期位置（Fig.2）では，PA2 のみを逃避者から遠ざけることで報酬の得やすさを変化させ，格差を実現している．実験環境における PA の状態は上下左右の壁，他の PA，逃避者の有無をもとに決定される．また，PA の行動は停止 or 上下左右への移動のいずれかとし，現在観測している状態が変化するまで同一の行動を取り続ける．1TP が経過したとする PA の行動選択回数は 500 回とした．逃避者の右隣と下隣に PA が同時に存在することを逃避者の捕獲とし，捕獲に関与した 2 体の PA に報酬が与えられる．報酬を獲得した PA の寿命は初期値（500）にリセットされ，1TP の間に報酬を獲得できなかった PA の寿命を 1 減少させる．

6 結果・考察

50TM を 1 実験として，従来設定および，改良後の設定において T_{GR}= 320 とした設定 A，T_{GR}= 1500 とした設定 B の 3 種類について，それぞれ 10 回の実験を行った．ここで設定 A では，20TP のうち 8 割程度の TP でいずれかの PA が報酬を獲得しつつ，全 PA が報酬を 1 回以上獲得している状況，

Table1 Comparison for each configuration

	従来設定	改良設定 A	改良設定 B
平均共存期間	14378.5TP	45541.9TP	43599.4TP
PA0,1 への平滑化行動選択率	17.3%	97.1%	97.4%
PA 報酬獲得率	51.0%	88.0%	93.4%

設定 B では 20TP の全てでいずれかの PA が報酬を獲得しつつ，全 PA が報酬を 4 回以上獲得している状況を想定し，T_{GR} を設定している．実験終期（41～50TM）の結果を Table1 に示す．ここで平滑化行動選択率は 50TM 終了時の値を，報酬獲得率については TM 終期（30000TP 以降）の値を掲載している．この表から，設定の改良によって PA 集団の共存期間を伸長できていることが分かる．また，PA0 と PA1 が報酬を独占している状態において，報酬を得難い PA2 へ報酬獲得を促す命令である「PA0,1 に対する平滑化行動」の選択率についても，設定の改良によって増加している．これは報酬獲得条件の変更によって問題(1)が解決され，適切な行動のみに報酬が与えられるようになったためと考えられる．さらに Table 1 より，平滑化行動の変更によって TM 終期における PA の報酬獲得率も大幅に増加していることが確認できる．これらの改良によって，多くの利益を獲得しながら格差解消・共存期間伸長を達成できたと考えられる．

また，設定 A,B では僅かながら結果に差が生じた．設定 A は MA の報酬獲得が容易であるためにルール重みが高値となり，PA 集団の共存に貢献しない行動の選択機会が減少するため，設定 B と比べて安定した共存を達成できた．しかし，設定 A の条件は PA の報酬獲得率が低い状況でも達成されるため，PA の報酬獲得率は設定 B の方が高値となった．PA の報酬獲得率の大きな減少は共存終了に繋がるため，報酬獲得条件の閾値が設定 A より低いと，共存期間が逆に縮小する事も想定される．加えて，命令選択間隔 20TP では MA が自身の状態を認識できない環境なども想定されるため，MA の報酬獲得条件の閾値 T_{GR} や命令選択間隔の適切な設定方法については，さらなる検討の余地がある．

7 まとめ

本稿では，相互作用型階層強化学習システム IH-RL に関する問題点を解決するため，MA の実験設定の改良を行った．計算機実験の結果，IH-RL を適用することで，追跡者エージェント集団の共存期間伸長を確認できた．今後は，環境規模や目的等の異なる様々な問題に対応できるよう，MA に関する効果的なパラメータ設定について検討していく必要がある．

参考文献

[1] 宮内龍之介，北越大輔，鈴木雅人：相互作用型階層強化学習システムを用いたエージェント群の協調行動獲得，第 19 回インテリジェント・システム・シンポジウム(FAN2009)講演論文集，pp484-489 (2009).

[2] 山﨑大地，北越大輔，鈴木雅人：相互作用型階層強化学習システムのマルチエージェント環境における特性評価，第 20 回インテリジェント・システム・シンポジウム(FAN2010)予稿集，S6-1-2 (2010).

[3] 今福啓：環境の変化に適応するマルチエージェントの学習手法，人工知能学会論文誌，Vol. 21，No. 2，pp. 153-166 (2006).