意図推定法を用いたマルチエージェント強化学習システムにおける協調行動の獲得

情報科学科　椿本 樹矢　　指導教員：小林 邦和

1　はじめに

Q 学習 [1] を用いたマルチエージェント環境下における協調動作に関して，他エージェントの内部モデルを知覚せず，行動のみの観測で協調を行えるモデルとして長行らの政策推定法が提案されている [2]．政策推定法は，他エージェントの行動一つひとつを予測し，それに合わせて自らの行動を決定するといった手法であるが，大容量の記憶が必須であるという問題も抱えている．

本研究では，少量の記憶容量を必要とし，設計者が意図的に協調を行えるよう報酬を操作しなくても，エージェント自らが考え，適切なゴールへ向かうための協調的な行動を獲得する手法を提案する．ここでは，2 体のエージェントが協力して複数ある重い荷物の中から一つを運ぶ場合を想定する．荷物を運ぶために，同じエージェントが同じ荷物の位置に到達することを協調とし，これを目的とする．

2　提案法

提案法では，他エージェントの行動一つひとつは最終的な目標を達成するためのプロセスであると考え，重要視しない．重要であるのは目標であり，それを推定することで自らの協調的な行動を獲得していく．

まず，自エージェントは時刻 t において他エージェント a の行動を観測し，図 1 に示すゴール g とのなす角度 θ，距離 d を用いて k の g への到達しやすさである優先度（GP : Goal Priority）を推定する．GP の計算式を式 (1) に示す．

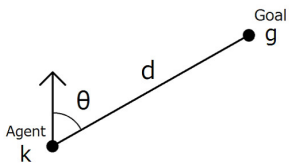

図 1　エージェントとゴールの位置関係

$$GP_{t,g,a} = (\alpha\, Dir_{t,g,a} + \beta\, Dis_{t,g,a} + Dir_{t,g,a}\, Dis_{t,g,a}) \,/\, (\alpha + \beta + 1), \tag{1}$$

$$\begin{cases} Dir_{t,g,k} = 0.5\,(\cos\theta + 1), \\ Dis_{t,g,k} = d^{-1}, \end{cases} \tag{2}$$

ここで，α, β は角度・距離をどの程度重要であると考えるかの割合である．特に，どちらが重要ということがなければ，通常はともに 1.0 でよい．

次に，算出した優先度をもとに，学習率 η を用いて自エージェントが考える各ゴールの価値である目標価値（GV : Goal Value）の更新を行う．

$$GV_{t,g} = (1-\eta)\, GV_{t-1,g} + \eta\, e(\overrightarrow{GP}_{t,g}), \tag{3}$$

$$\overrightarrow{GP}_{t,g} = (GP_{t,g,1}, GP_{t,g,2}, \cdots, GP_{t,g,n}), \tag{4}$$

ここで，n は他エージェントの数，e() は任意の評価関数であり，エージェントが行うべき協調に合わせた関数を設定する．また，$\overrightarrow{GP}_{t,g}$ は自分以外の全ての $GP_{t,g,k}$ をもつベクトルである．この目標価値 $GV_{t,g}$ は他エージェントの行動から推定した意図を反映したものとなっている．

これらの値は毎ステップ更新され，自エージェントのゴール到達時に利用される．時刻 t に自エージェントがゴール g へ到達した時，環境から報酬 r_g が与えられると同時に，目標価値を用いて，報酬の解釈を式 (5) で行う．

$$r_t \leftarrow r_t \times 2\, GV_{t,g}. \tag{5}$$

その後，解釈された報酬が Q 値の更新式に取り入れられ，協調行動を獲得していく．

3　計算機シミュレーション

2 次元 (20 × 20) グリッド空間中に，2 体のエージェントと 2 つのゴールが存在する環境を用いる．周囲は壁に囲まれており，エージェントとゴールはランダムに配置される．全てのエージェントが同じゴールに到達することを協調の目的とする．同じゴールに到達したエピソードを成功エピソードとし，各エピソードでの成功確率，つまり協調の成功率を評価基準とし，エピソードと成功率の推移を比較する．グラフは 100 回のシミュレーションの平均値である．評価関数 e() は平均 E[] とした．

通常の Q 学習，協調を用いた Q 学習，提案法 (Q-learning with Intention Estimation) の成功回数の推移を図 2 に示す．

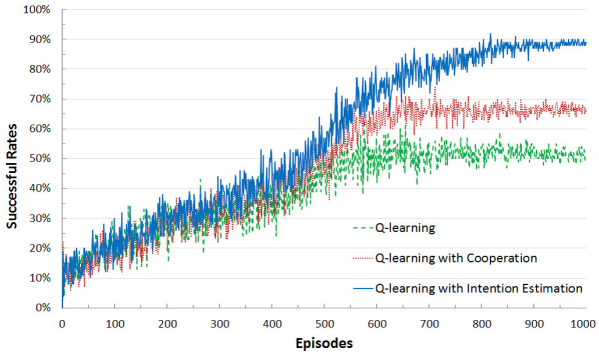

図 2　成功率の推移

提案手法，協調を用いた Q 学習が協調行動の獲得に成功していることがわかる．通常の Q 学習は学習後期においても成功回数が増加することはなく，協調的な行動を獲得できていない．また，協調を用いた Q 学習より，提案手法の方が成功回数の面で優れていることがわかる．

4　まとめ

本稿では，協調動作を獲得する新たな手法を提案した．また，評価シミュレーションにより協調動作の確認を行った．

参考文献

[1] R. S. Sutton, A. G. Barto: 「強化学習」, 森北出版 (2000)
[2] 長行 康男, 伊藤 実: 電子情報通信学会論文誌 D, Vol. J86-D1, No. 11, pp. 821-829(2003)
[3] Tatsuya Tsubakimoto, Kunikazu Kobayashi: Proc.of ICAROB 2014, pp.122–125 (2014)