行動プランの交渉を伴うマルチエージェント学習に関する考察

河目　瞬*　吉川　毅†　栗原　正仁‡
北大　北大　北大

1 はじめに

エージェントが複数存在する環境では，各エージェントにとって，他のエージェントの動きを知覚することは重要である．その理由は，エージェントは環境を通して互いに影響し合い，それが各エージェントの行動決定の要因に成り得るためである．このようなマルチエージェント環境においては，相手の行動に合わせて自分の戦略を決定することは重要だが，もし，自分の戦略に相手の行動を合わせることができれば，より自分に有益な結果を得ることができると思われる．さらに，マルチエージェント学習において，各エージェントに「協調する行動」を学習させようとした場合，エージェント同士の大域目的が同じであれば，個々が独立に学習していたとしても，いずれは協調するような行動を学習する場合がある．しかし，各エージェントの目的が異なる場合に，他のエージェントの協力が得られるような行動を個々の学習で獲得するのは難しい．この場合，他のエージェントに協力を依頼することが，現実問題においても妥当な考えである．

本研究では，マルチエージェント学習の過程に交渉過程を含めたエージェントのモデルを提案する．エージェント間で相手の行動を要請，そしてその受容／拒否をすることができるエージェントを構築し，強化学習の枠組みを利用した学習方法に適用させることを目的とする．また，目標の異なるエージェント同士に，協調を学習させる問題に対してシミュレーションを行うことにより、その有効性を確かめる．

Study for the Multi-Agent Learning Through a Negotiation

*Shun KAWAME, Hokkaido University
†Takeshi YOSHIKAWA, Hokkaido University
‡Masahito KURIHARA, Hokkaido University

2 強化学習の枠組み

強化学習は，エージェントが環境に対して試行錯誤した結果環境から得られる，強化信号（報酬）のみを用いて環境に適応していく学習制御の枠組みである．強化学習はその性質上，不確実性のある環境を扱っている．学習を行うエージェントが複数いるようなマルチエージェント環境では，自分以外のエージェントの取る行動は不確実性を持っており，強化学習の枠組みに当てはまる。

さらに強化学習では，報酬に遅れが存在するような段階的な行動の順序の獲得も扱っている．エージェント同士の協調を学習させる際，エージェントの行動に対する即時的な評価のみで学習をさせるのは得策ではない．その理由は，協調行動の獲得のためには，妥協と呼ばれる行動も必要なためである．よって，本研究では，エージェント同士の関係の学習に，強化学習の枠組みを適用することにする．

本研究では、強化学習の一般的な手法の一つである Q 学習を採用する．Q 学習では，システム状態の集合 Σ と，可能な行動集合 A を考慮する．エージェントが状態 $S_t \in \Sigma$ において，行動 $a_t \in A$ を取るとき，報酬 r_t が与えられ，状態 S_{t+1} へ推移する．ここで，状態 S において行動 a を取ることで得られる報酬の期待値の総計として Q 値を定義し，この Q 値を以下の更新式 (1) で更新していく学習方法が，Q 学習である．

$$\begin{aligned} Q(S_k, a_k) = (1-\alpha)Q(S_k, a_k) \\ + \alpha(r_k + \gamma \max_a Q(S_{k+1}, a)) \end{aligned} \quad (1)$$

ただし，γ（$0 < \gamma < 1$）は，無限の総和の保証をし，将来の報酬はより少ないと評価する割引率で，α（$0 < \alpha < 1$）は，学習率である．

3 エージェントモデル

エージェントは，強化学習の枠組みを用いて最適な行動選択法（方策）を学習していくが，本研究

ではまず，この行動群の中に交渉するという選択肢および相手の交渉を受諾するか拒否するかの選択肢を加える．そして，Q 学習で用いられている Q 値の概念を利用し，そのエージェント間による交渉の有意義さを表す指標を導入する．エージェントには試行錯誤の結果，この指標を獲得させていき，それにより，相手エージェントに合わせた単独行動と，相手エージェントとの交渉を通じた協調行動のどちらがより有益かを学習させることを実現させる．

2 人のエージェントを A および B で表した時，エージェント A の取り得る行動を $a = \{a_1, a_2, \ldots, N(B), A(B), D(B)\}$ とする．$N(B)$ はエージェント B と交渉を行うという行動を示しており，B に対して B の行動群を提示する．$A(B)$ はエージェント B の交渉を受容するという行動を示しており、B から提示された行動群を実行する．$D(B)$ はエージェント B の交渉を拒否するという行動を示しており、自身の行動選択法に基づいた行動を実行する．各エージェントは，各状態行動対における Q 値を持つ．さらに、A と B の間の交渉の有意義さを表す指標として，次式 (2) で表される値，Q_N も持つ．

$$Q_N = f(N_r, (Q(S, N(B)), Q(S', A(B))) \quad (2)$$

ただし，N_r は交渉成功率である．エージェントは，この Q_N と，単独行動における Q 値とを比較して行動選択をする．

4 実験設定

本稿では，異なる目標を持つエージェント同士による協調の学習の簡単な問題として，次の問題を考える．

実験で扱うエージェントを $\{A, B\}$ で表し，各エージェントの知覚できる状態を $S = \{S_0, S_1\}$ とする．各エージェントの取り得る行動は，エージェント A が $a = \{a_1, a_2, N(B), A(B), D(B)\}$ エージェント B が $b = \{b_1, b_2, N(A), A(A), D(A)\}$ である．ただし、状態 S_0 においては、エージェント A のみが交渉を持ちかける行動 $N()$ を行使でき、逆に状態 S_1 においては、エージェント B のみが交渉を持ちかける行動 $N()$ を行使できるものとする．それぞれの状態における各エージェントの利得行列を図 1 のように設定する．

状態は $S_0 \rightarrow S_1 \rightarrow S_0 \rightarrow \ldots$ と，エージェントの取った行動に依らず 2 つの状態を繰り返し遷移

S_0

A	b_1	b_2
a_1	0	-2
a_2	-7	10

B	b_1	b_2
a_1	-2	0
a_2	2	-4

S_1

A	b_1	b_2
a_1	-2	2
a_2	0	-4

B	b_1	b_2
a_1	0	-7
a_2	-2	10

図 1: 利得行列

するものとし，$S_0 \rightarrow S_1$ の区切りで 1 ターンと呼ぶこととする．また、エージェントの取り得る行動 $N(B)$ および $N(A)$ は，両者間の可能な行動の組み合わせの中から，自分が有益になるような行動対を探索し，その行動を相手に提示する行動である．今回の実験では、これを $a_2 \times b_2$ と固定する．すなわち，相手に提示する行動は A の場合 b_2，B の場合 a_2 となる．

この問題では，エージェントが両者間でコミュニケーションを持たずに，仮にミニマックス法の戦略を両者が取ったとすると，S_0 では $a_1 \times b_1$ に，S_1 でも $a_1 \times b_1$ へと行動対が決まり，1 ターンあたりの獲得利得は，$0 + (-2) = -2$ となる．

この問題を，交渉を行うエージェントモデルのものと，交渉を行わないエージェントモデルのものとで実験し，結果を比較する．

5 おわりに

本研究では，マルチエージェント学習の過程に交渉過程を含めたエージェントのモデルを提案した．また、目標の異なるエージェント同士の協調の学習を目的とした問題に対して、このモデルのエージェントを適用し、その有効性を確かめた．

参考文献

[1] Richard S. Sutton and Andrew G.Barto
三上 貞芳，皆川 雅章　共訳: “強化学習”
森北出版，2000 年

[2] 高玉　圭樹　著: “マルチエージェント学習”
コロナ社，2003 年