E-018

非ゼロ和ゲームにおける混合戦略の強化学習 -１・２・５じゃんけんを例に-

Reinforcement learning of mixed strategies for non-zero-sum game - 1-2-5 janken as an example -

後藤 強 † 伊藤 昭 ‡ 寺田和憲 ‡
Tsutomu Goto Akira Ito Kazunori Terada

1. はじめに

我々は、知的なエージェントがゲーム理論的状況で、どのようにして相手の行動を読み、自己の最適行動を生成するのかを研究している。そして、我々は Q 学習に履歴というものをを導入することにより、繰り返し対称非ゼロ和ゲームに Q 学習が適応できることを示した [5]。また、この履歴を用いた Q 学習は、お互いに話し合うことによって協調するのではなく、独立して行動した結果として、協調することを発見し、高い得点を得ることができた。

しかし、我々はこの履歴を用いた Q 学習が最も良い戦略ではないと考えている。なぜならば、Q 学習は今までに得られた報酬から、将来最も多い報酬を得られるような行動を学習しているからである。また、Q 学習はゲーム理論的状況で有効とされている、混合戦略を学習することができないからである。

我々は、混合戦略を学習する方法として、MiniMaxQ 学習 (MQ) と誘導型 Q 学習 (IQ) を用い、また非ゼロ和のゲーム理論的問題の例として１・２・５じゃんけんを用いて、それぞれの戦略の性質を調べた。

2. １・２・５じゃんけん

我々がゲーム理論的問題の例として用いた１・２・５じゃんけんについて説明する。この１・２・５じゃんけんというゲームは、普通のじゃんけんと違い、勝ち負けだけでなく、どの手で相手に勝ったかによって得られる得点が違うというゲームである。つまり、グー (G) で勝つと１点、チョキ (C) で勝つと２点、パー (P) で勝つと５点の得点を得られるというゲームである。

１・２・５じゃんけんの利得表を以下の表１に示す。

	G	C	P
G	0,0	1,0	0,5
C	0,1	0,0	2,0
P	5,0	0,2	0,0

表 1: １・２・５じゃんけんの利得表

このゲームでは、どのような戦略が考えられるだろうか。明らかに特定の手を出す純戦略は最適戦略になり得ない。しかしながら混合戦略の範囲では、互いに相手の手の最適戦略になっている Nash 均衡解が必ず存在する。このゲームでは Nash 均衡解は双方が G,C,P を確率 $(x,y,z)_N = (2/17, 10/17, 5/17)$ で出すことで、そのときの平均利得は 10/17 となる。また、相手の手に関わらず自己の最低点を確保する手 (MinMax 戦略) としては、$(x,y,z)_M = (10/17, 5/17, 2/17)$ が存在し、そのときの平均利得はやはり 10/17 である。一方、両者がランダムに手を選んだ場合 ($(x,y,z)_R = (1/3, 1/3, 1/3)$) では、お互いに得られる平均利得は、8/9 となる。

これらは 1 回限りの対戦における戦略である。しかし、この１・２・５じゃんけんを繰り返し行う場合では、異なった戦略がある。例えば、お互いに G と P を出し、交代で P で勝つというような協調行動をとることにより、5/2 という高い平均得点を得ることができる。

3. 履歴を用いた Q 学習

Q学習とは、過去に状況 S で自分が選択した行動 a とその時得られた報酬 r の組から、現在の状況 S_t で行動 a_t をとるときの割引された報酬の和の期待値 $Q(S_t, a_t)$ を計算し、行動を選択する。しかしながら、相手の戦略が決まらない (分からない) 限り自己の報酬の期待値は求まらない。そこで、相手は過去 h 回の双方の手の組によって次の出す手を決めている「h 次マルコフ戦略」と仮定すると、過去 h 回の相手と自分の手の組

$\{a^m_\tau, a^o_\tau\}^{t-h}_{t-1} = \{a^m_{t-1}, a^o_{t-1}, a^m_{t-2}, a^o_{t-2}, \cdots, a^m_{t-h}, a^o_{t-h}\}$

を状態と考えることで、相手を含む世界を MDP(マルコフ決定過程) としてモデル化することができる。

状態 S_t で自分が a^m、相手が a^o をとり、報酬 $r(a^m, a^o)$ を得て、状態 S_{t+1} に移行したとき、Q 値の更新は、以下の式で行う。

$$Q(S_t, a^m) \leftarrow (1-\alpha)Q(S_t, a^m) + \alpha(r(a^m, a^o) + \gamma \max_{a^m} Q(S_{t+1}, a^m))$$

行動選択は、ϵ-greedy と Boltzmann 型とを組み合わせて行う。すなわち、以下アルゴリズムで行動を決定している。

・ϵ の確率でランダムに行動 a^m を選択
・それ以外
　$p(a^m) = C\exp(Q(S, a^m)/T)$ の確率で行動 a^m を選択

4. MiniMaxQ 学習と誘導型 Q 学習

MiniMaxQ 学習 (MQ) とは、ゲーム理論の MiniMax 原理を強化学習に取り込んだものである。すなわち、相手が自分の報酬を最小にしようと行動を選択すると仮定し、自分の最低点を確保するような行動を選択するというものである。

状態 S_t で自分が a^m、相手が a^o をとり、報酬 $r(a^m, a^o)$ を得て、状態 S_{t+1} に移行したとき、MiniMaxQ 学習で

†岐阜大学大学院工学研究科
‡岐阜大学工学部

の Q 値の更新式は以下の式になる。

$$Q(S_t, a^m, a^o) = (1-\alpha)Q(S_t, a^m, a^o) + \alpha(r + \gamma V(s_{t+1}))$$

行動確率 $p(S_t, a)$ は、線形計画法を用いて以下の式で計算する。

$$p(S_t, a) = \arg\max\{p(S_t, a), \min_{a^o}(\sum_{a'} pi(S_t, a')Q(S_t, a', a^o))\}$$

また、ϵ-greedy 型の行動選択も用いている。すなわち、
・確率 ϵ でランダムに行動を選択
・それ以外は、$p(S_t, a)$ の確率に基づいて行動を選択
また、Q 値を計算するときに用いた行動価値 $V(S_t)$ の計算は以下の式で行う。

$$V(S_t) = \min_{a^o} \sum_{a'} pi(S_t, a')Q(S_t, a', a^o)$$

誘導型 Q 学習 (IQ) とは、MiniMaxQ 学習と同じように、混合戦略を学習する方法である。

1. 自分を混合戦略と仮定する
2. 相手を Q 学習と仮定する
3. 相手の Q 値を自分の得点が高くなるよう誘導するために、自分の行動確率 $P(S_t, a)$ を変化させる
4. 行動選択は ϵ-greedy 型を組み合わせて、以下のように決定する
   ・確率 ϵ でランダムに行動を選択
   ・それ以外は、$P(S_t, a)$ の確率で行動を選択

5. 実験結果

各エージェントの振る舞いを調べるため、実際に各エージェント同士を対戦させた。実験に用いたエージェントは、履歴を使用する SQ・MQ・IQ, また、Nash 均衡解の確率 $(G, C, P) = (2/17, 10/17, 5/17)$ で手を選択する Nash, GCP を等確率に選択する Random を用いた。

学習に用いたパラメーターは $\alpha = 0.1$, $T = 0.2$, $\gamma = 0.9$, $\epsilon = 0.01$ とし、また Q の初期値は 0 とした。各対戦は 10^8 回のじゃんけんを 1 試合とし、その最後の 10^7 を収束値と考え平均値を求める。これを乱数を変えて 10 回行い平均を取った。表の数値は、左側の戦略が上側の戦略と対戦したときの、左側の戦略の平均得点である。

表 2 は履歴長が 2 の場合における対戦結果を示している。

	Random	Nash	SQ2	MQ2	IQ2
Random	0.889	0.591	0.671	1.154	0.567
Nash	0.921	0.592	0.945	1.038	0.486
SQ2	1.610	0.590	2.298	3.43	1.984
MQ2	0.591	0.588	0.210	0.550	0.560
IQ2	0.778	0.592	0.404	0.952	0.585

表 2: 対戦結果

表 2 から、以下のことがわかる。
・履歴を用いた SQ は、相手に合わせて行動し、高い平均得点を得ている。
・履歴を用いた SQ 同士では、協調行動をすることでお互いに高い平均得点をあげている。
・MQ と IQ は、あまり得点が高くない。
・Nash は、対戦した相手の得点を約 0.59 にする。
・ほとんどのエージェントが、Nash との対戦で相手よりも低い得点をとっているのに対して、逆に IQ は Nash より高い得点を得ている。
・MQ は、相手が最悪の行動を選択すると仮定して、自身の行動を決定しているため、平均得点は低くなる。

IQ は、他のエージェントとの対戦結果はあまり良いとは言えない。しかし、Nash と対戦した場合では、Nash より良い得点をあげている。

IQ 型戦略の意味は次のようなものである。仮に、相手が Nash 均衡解の確率で手を選択するという戦略を学習した学習エージェントだとする。もし相手が、このまま同じ戦略をとり続けるのならば、自分の得点は上がることはない。では、自分の得点を上げる為にはどうしたら良いのだろうか。相手に Nash 戦略を止めさせ、別の戦略を取らせられれば、自分の得点を上げることができる。そのためには、相手がこのまま同じ戦略をとり続けるのならば、得点が下がるという事を自身の行動で示すことである。

今回の場合では、Nash は固定戦略であるため、誘導型 Q 学習 (IQ) を用いても戦略を変更することはない。しかし、学習エージェントを含む様々な戦略と対戦した場合に、相手に戦略の変更を迫る (誘導する) といったような能力を持つというようなことは、必要になる時があると我々は考える。

6. パラメーター P_{th} の導入

MQ も IQ もあまり得点が高くない。これは, 相手が行動を選択する可能性が極めて低いものまで考慮し、自身の行動を決定している為だと考えられる。すなわち、Q 値から得られる最善の行動でも、仮に、相手が選択する可能性がほとんど無いのならば、自分の行動を選択するときに考慮する必要はないのではないだろうか。

そこで、新しいパラメーター P_{th} を導入する。これは、各状態における相手の行動選択の確率を記憶しておき、この行動確率が P_{th} 以下になったら、それに対応する Q 値を 0 にし、自分の行動選択時に考慮しないようにする。これによって、行動選択時の探索範囲を制限することで、性能が向上すると考える。

7. 実験結果

導入したパラメーター P_{th} の効果を調べるため、再びリーグ戦を行った。使用したエージェントは、前回と同じである。履歴を使用する SQ・MQ・IQ と Nash・Random である。

学習に用いたパラメーターは $\alpha = 0.1$, $T = 0.2$, $\gamma = 0.9$, $\epsilon = 0.01$,$P_{th} = 0.1$ とし、また Q の初期値は 0 とした。各対戦は 10^8 回のじゃんけんを 1 試合とし、その最後の 10^7 を収束値と考え平均値を求める。これを乱数を変えて 10 回行い平均を取った。表の数値は、左側の戦

略が上側の戦略と対戦したときの、左側の戦略の平均得点である。

	Random	Nash	SQ2	MQ2	IQ2
Random	0.889	0.591	0.670	1.152	0.571
Nash	0.921	0.592	0.945	1.038	0.386
SQ2	1.610	0.590	2.298	1.573	2.279
MQ2	0.593	0.588	1.891	1.344	0.732
IQ2	0.795	0.590	2.068	1.175	0.678

表 3: P_{th} を導入した場合の対戦結果

表 3 より、閾値 P_{th} を導入することにより、MQ、IQ ともに得点が上昇していることがわかる。IQ では、Nash と対戦した場合、P_{th} を導入しても、相手よりも高い得点を取っている。また、ゲーム理論的強化学習 (MQ・IQ) でも、いくつかの対戦では協調行動が見られた。

次に、この閾値 P_{th} の効果を調べるために、MQ の P_{th} の値を変え、対戦させた。対戦結果を図 1 に示す。対戦に用いたエージェントは、SQ、閾値 P_{th} を用いない MQ、$P_{th} = 0.1$ の MQ である。各エージェントの履歴長は 2 である。

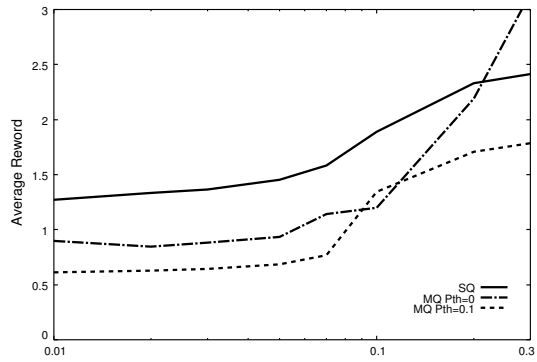

図 1: P_{th} を変化させた時の平均得点の変化

図 1 より、P_{th} を大きくした場合、MQ は SQ と同じような振る舞いを見せるようになり、協調行動が起きているのではないかと考えられる。

SQ-MQ の対戦における各戦略の平均得点の時間変化を図 2 に示す。また、SQ 同士の対戦における各戦略の平均得点の時間変化を図 3 に示す。履歴長は SQ・MQ ともに 2 であり、MQ のパラメーター $P_{th} = 0.3$ である。

図が 2.5 付近で振動しているのは、確率 ϵ で行動をランダムに決定するためである。

学習速度の違いはあるが、最終的にはどちらの学習方法も協調行動により高い平均得点をあげている。

8. まとめ

ゲーム理論的強化学習である、MiniMaxQ 学習 (MQ) や誘導型 Q 学習 (IQ) に閾値 P_{th} を導入したことにより、協調行動が発生し平均得点が上がった。だが、すべての状況で良い得点を得られるわけではない。

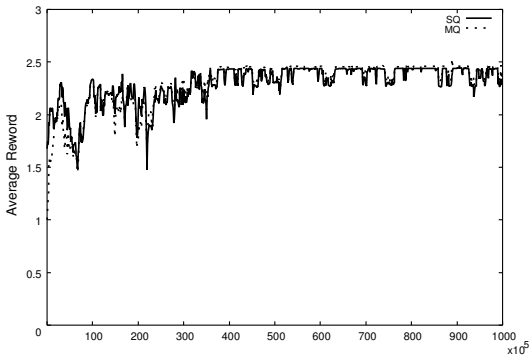

図 2: 平均得点の時間変化 (SQ-MQ)

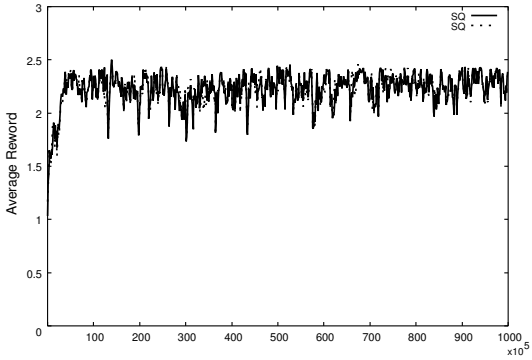

図 3: 平均得点の時間変化 (SQ-SQ)

そのため、相手に応じて学習方法を切り替えるといったような方法も必要ではないだろうか。また、使用する履歴の長さについても、エージェント自身が決定しているわけではない。履歴の長さを自動的に決定するような方法を考えることも今後の課題である。

参考文献

[1] Sutton, R.S. and Barto, A.G.: Reinforcement Learning, The MIT Press, 1998.

[2] Littman, M. L.: “Markov games as a framework for multi-agent reinforcement Learning,” 11th International Conf. on Machine Learning, ML-94, pp.157-163, 1994

[3] Hu, J. and Wellman M.P.: “Multiagent reinforcement learning: Theoretical Framework and an Algorithm,” 15th International Conf. on Machine Learning, ML-98, pp.,1998

[4] 伊藤昭: “心を読む能力の創発 -マルチプレーヤー囚人のジレンマゲーム-” , 認知科学,Vol.6 No.2,1999.

[5] 後藤 強, 岩田 元志, 伊藤 昭, 寺田 和憲: “繰り返し対称非零和ゲームの強化学習 -1・2・5・じゃんけんを例に-” 情報処理学会 第 66 回全国大会