ISS-P-68

学習による協力関係構築に関する１・２・５じゃんけんの考察

仁科　諒子†　　竹川　高志†
†工学院大学情報学部情報デザイン学科

1. はじめに

非ゼロ和ゲームの「1・2・5じゃんけん」では，相互に協調することで高得点を得られることがわかっている[1]．しかし選択された手を見ただけではプレイヤーの明確な意図がわからず，また協調といえる状態が複数ステップに及ぶため協調が難しい．

本研究では「1・2・5じゃんけん」を用いて，強化学習プレイヤーに行動戦略を学習させる．プレイヤーが対戦相手と自分の行動履歴に基づき，対戦を行うことでQ学習によって行動戦略を学習する様子をシミュレーションし，人間社会においてどのように行動すればお互いに協調し合い社会全体の利益を増やすことができるか考察を行った．

2. 1・2・5じゃんけん

1・2・5じゃんけんとは，通常のじゃんけんと同じ「グー(G)」，「チョキ(C)」，「パー(P)」で勝敗を決めるが，勝ったときの手により得点が変わるゲームである．

グー・チョキ	グー選択側は+1 点 チョキ選択側は−3 点
チョキ・パー	チョキ選択側は+2 点 パー選択側は−3 点
パー・グー	パー選択側は+5 点 グー選択側は−3 点
あいこ	各々加点なし

両者がG，C，Pを等確率で選んだ場合は，お互いに得られる平均利得が−1/9となる．この得点構造の場合，プレイヤー2 人が協調し交互に G と P を出し合う戦略は，5 点と−3 点を交互に得ることで安定して得点を増やすことができる．両者が始めから協調できた場合，お互いに得られる平均利得は，1 となる．

3. 方法

実験では，学習しないプレイヤーとして対戦相手に「歩み寄り戦略(CMP)」「GP 交互戦略(ALT)」「GP 等確率戦略(GPR)」「等確率戦略(EQP)」「常にP戦略(POL)」を用いた．

CMP は「囚人のジレンマ」における「しっぺ返し戦略」に相当し，相手が協調していると思われるなら自分も協調を続け，逆に相手が協調する様子を見せないならば自分も協調しない(表 1)．状態 s を過去 2 手の対戦履歴，行動 a を出す手として Q 学習を行った．

表1：歩み寄り戦略

最初の 4 行はそれぞれが対戦時に出した手の履歴である．CMP(A)，CMP(B)の列は選択する手である．R は GCP の中から 1/3 の確率で手を選択し，GP は GP の中から 1/2 の確率で手を選択する．

相手の1手前	G			C	P											
相手の2手前	G	C	P		G			C			P					
自分の1手前					G	C	P	G	C	P	G	C	P			
自分の2手前													G	C	P	
CMP(A)	GP	R	G	R	P			R			GP					
CMP(B)	GP	R	G	R	P	C	P	R	C	R	GP		C		GP	

4. 実験

学習結果は 5 つの戦略のプレイヤーとの対戦結果で評価した．単独の単純戦略で学習した場合，全ての場合において相手の戦略の攻略法を見つけた．例えば，CMP，GPR で学習した Q 学習プレイヤーは POL に対して P を出し続けた．GPR と対戦した Q 学習プレイヤーは，CMP(A)，ALT，GPR，POL に対して P を出し続けた．未来報酬の割引率γが高いとき，CMP(B)で学習した Q 学習プレイヤーは，CMP に対して GPP の順で手を出し続けた．EQP と対戦した Q 学習プレイヤーは ALT に対して，γが低いときは CPPP の順で手を出し続けた．

5. まとめ

今回の条件のように個々の単純な戦略で学習させた場合，Q 学習の戦略はパターンを読みきり，協調より絶対的な勝利に推移する．人間社会のように対戦相手が 1 人に限らず，またその戦略もさまざまである場合の学習で得た戦略を評価することが今後の課題である．

参考文献

[1] 伊藤昭，大橋資紀，寺田和憲：「非零和ゲームの強化学習−相手の行動を読むプログラム」，情報処理学会研究報告．ICS，[知能と複雑系]，53-60，2005