5R-6

Profit Sharing を用いたぷよぷよの学習

小林港　長名優子

東京工科大学大学院 バイオ・情報メディア研究科コンピュータサイエンス専攻

1 はじめに

強化学習 [1] に関する研究としてゲームの戦略獲得を題材とした研究が広く行われている．しかし，これらの研究の多くではエージェントを含む環境は完全観測であると仮定されていたり，部分観測環境を対象としたものであったとしても比較的小さい状態空間で表現されるゲームのみを対象としていたりと様々な制約がある．

本研究では，Profit Sharing[2] を用いて落ち物パズルゲームのひとつであるぷよぷよの学習を行う．Profit Sharing のような強化学習では一般に試行錯誤を繰り返すことで報酬を得るためのルールを獲得していくが，本研究では，人間が実行したプレイデータをエピソードとして用いることで学習を行う．また，Profit Sharing により獲得したルールを解析することで，学習に用いるプレイデータの違いにより学習されるスキルの違いがみられることなどを調べる．

2 Profit Sharing

Profit Sharing[2] では，エージェントは報酬が得られるまで行動を行い，エピソード内の各ルールに報酬を分配する．ここで，エピソードとはエージェントが行動を開始してから報酬を得るまでのルール系列をさす．また，ルールは状態 s と行動 a の対であり，(s,a) で表される．Profit Sharing では一般には報酬が得られた時刻から遠い時刻のルールほど分配される報酬が少なくなるように設定される．時刻 t におけるルールに報酬をどのように分配するかを決める強化関数 $F(t)$ を

$$|C^A|\sum_{i=1}^{t-1}F(i) < F(t) \tag{1}$$

を満たすように設定することでマルコフ決定過程においては合理的政策を獲得できることが知られている [3]．ここで，合理的政策とは単位行動あたりの期待報酬獲得量が正であるような政策をさす．式 (1) において $|C^A|$ は同一の状態においてとり得る行動の種類の数を表す．式 (1) の条件を満たす強化関数として

$$F(t) = \frac{1}{|C^A|^{W-t}} \tag{2}$$

のような等比減少関数が一般に用いられる．ここで，W はエピソードの長さを表している．

Profit Sharing では，エージェントはある状態におけるルールの価値の比に基づいて行動を決定する．行動の決定には，一般にルーレット選択やボルツマン選択などが用いられる．ルーレット選択において状態 s のときに行動 a を選択する確率 $P(s,a)$ は

$$P(s,a) = \frac{q(s,a)}{\sum_{b\in C^A} q(s,b)} \tag{3}$$

で与えられる．また，ボルツマン選択において状態 s のときに行動 a を選択する確率 $P(s,a)$ は

$$P(s,a) = \frac{\exp(q(s,a)/T)}{\sum_{b\in C^A} \exp(q(s,b)/T)} \tag{4}$$

で与えられる．ここで，$q(s,a)$ は状態 s において行動 a をとるというルール (s,a) の価値であり，ルール (s,a) が報酬獲得にどのくらい貢献するかを表している．また，C^A はエージェントのとり得る行動の集合，T は温度パラメータである．Profit Sharing では，エージェントは報酬が得られるまで行動を行い，エピソードごとにルールの価値の更新を行う．ルールの価値は，報酬を分配することで

$$q(s_t,a_t) \leftarrow q(s_t,a_t) + r\cdot F(t) \tag{5}$$

のように更新される．ここで，s_t は時刻 t における状態，a_t は時刻 t における行動であり，t はエージェントが行動を開始してからの時刻を表している．

3 Profit Sharing を用いたぷよぷよの学習

3.1 ぷよぷよ

ぷよぷよは落ち物パズルゲームのひとつであり，一般には 2 人で行う対戦型のアクションパズルゲーム

Learning on Puyopuyo by Profit Sharing
Minato Kobayashi and Yuko Osana (Tokyo University of Technology, minato@osn.cs.teu.ac.jp, osana@cs.teu.ac.jp)

として知られている．本研究では，ぷよぷよを対戦型ゲームとしてではなく，1人で行うパズルゲームとしてとらえ，学習を行う．

ぷよぷよでは，横6マス × 縦12マスの格子上のフィールドに2つ1組で落下してくる「ぷよ」と呼ばれるブロックを操作し，同じ色のぷよを4つ以上連結させることで消すことを目的としてプレイする．なお，ぷよの色は本研究では4色としている．ぷよぷよでは，同じ色のぷよが4つ以上連結して消滅することで得点が得られ，ぷよの消滅により上にあったぷよが落下することで4つ以上のぷよが連結して消滅する連鎖が発生するとその連鎖の回数に応じて得点が加算されることになる．

ぷよぷよに関してはフィールド上のぷよをすべて消せるかどうかを判定する全消し判定問題 [4] やフィールドの状態から連鎖数を判定する問題 [5] などに関する研究が行われており，いずれの問題もNP完全であることが知られている．

3.2 ぷよぷよの学習

Profit Sharingを用いたぷよぷよの学習において，すでにフィールドに積まれているぷよの配置を状態，次に落ちてくる2つ1組のぷよをどこにどのような向きで配置するかを行動とし，これらの対をルールとする．ぷよを消して得点が得られるまでを1つのエピソードとして扱い，得られた得点に基づく報酬をエピソードに含まれるルールに分配することで学習を行う．強化学習では一般に試行錯誤を繰り返すことで報酬を得るためのルールを獲得していくが，ここでは人間が実行したプレイデータをエピソードとして用いることで学習を行う．

ぷよぷよの学習において，フィールドに積まれているぷよの配置を状態として扱うが，ゲームが進むにつれて状態数が増えてしまい計算が困難になるという問題がある．また，ゲーム中に出現する様々な状態に対して適切な行動を学習するには膨大な量のプレイデータが必要となる可能性がある．したがって，フィールドのぷよの配置をそのまま状態として扱うことは望ましくないと考えられる．本研究では，次に落ちてくるぷよの配置の仕方を決定する際に考慮するのは配置しようとするぷよの色と同じ色のぷよの配置であると考え，状態の表現を行う．例えば，図1において(a)〜(e)の状態は別の状態であるが，配置しようとするぷよの色と同じ色のぷよの配置のみに着目して表現するとすべて(f)のように表現することができ，同じ状態として扱うことが可能になる．なお，図1(a)〜(e)において○，□，△，×は4色のぷよを表している．また，図1(f)において★と☆は配置しようとするぷよと同じ色のぷよ，※はそれ以外の色のぷよを表している．なお，(b), (c)では落下中のぷよの色の上下が(f)とは逆になっているが落下中のぷよは配置する際に回転することが可能なため，同じ状態として扱うことができる．

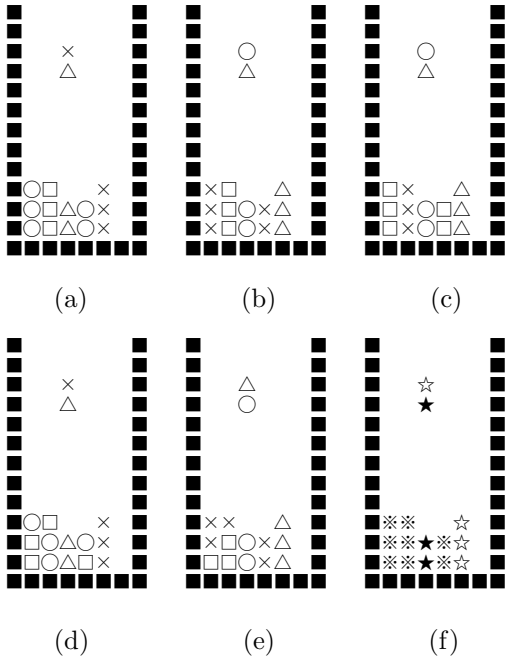

図1: ぷよぷよにおける状態の表現

4 計算機実験

提案手法を用いて計算機実験を行い，ぷよを消すことができるような戦略を獲得できることを確認した．

参考文献

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

[2] J. J. Grefenstette : “Credit assignment in rule discovery systems based on genetic algorithms,” Machine Learning, Vol.3, pp.225–245, 1988.

[3] 宮崎和光, 山本雅幸, 小林重信 : “強化学習における報酬割り当ての理論的考察,” 人工知能学会誌, Vol.9, No.4, pp.580–587, 2008.

[4] 牟田秀俊 : “ぷよぷよはNP完全,” 電子情報通信学会技術研究報告, COMP105-72, pp.39–44, 2005.

[5] 松金輝久, 武永康彦 : “組合せ最適化問題としてのぷよぷよの連鎖数判定問題,” 電子情報通信学会論文誌 D, Vol.J89-D, No.3, pp.405–413, 2006.