ドットイートゲームへの Q 学習の適用
Applying Q-Learning to a Dot-Eat Game

森田　隆弘
Takahiro Morita
法政大学情報科学部コンピュータ科学科
E-mail: takahiro.morita.4p@stu.hosei.ac.jp

Abstract

This paper proposes UCB fuzzy Q-learning by combining fuzzy Q-learning and the UCBQ algorithm and applies it to a dot-eat game. The UCBQ algorithm improved the action selection method called the UCB algorithm by applying it to Q-learning. The UCB algorithm selects the action with the highest value, called the UCB value, instead of a value estimate. In addition, since this algorithm is based on the premise that any unselected actions are selected and value estimates are obtained, the number of unselected actions is small, and it is possible to prevent local solutions. The proposed method aims to promote learning efficiently by reducing unselected actions and preventing the Q value from becoming a local solution in fuzzy Q-learning in a dot-eat game. This paper applies the proposed method to a dot-eat game called Ms. PacMan, and presents an experiment on finding optimum values used in the method. Its evaluation is performed by comparing the game score with the score obtained by the AI of a previous study. The result shows that the proposed method significantly reduced unselected actions.

1. はじめに

近年，AI に高い関心が集まっており，ゲーム AI の研究も盛んに行われている．2017 年にはプロ棋士に勝利した将棋 AI・Ponanza が現れ，ゲーム AI が多くの人に知られるようになった．さらに囲碁でも，同年にプロと囲碁 AI・AlphaGo の三番勝負で，AI が 3 局全勝をあげたことで大きな話題となった．ゲーム AI の多くで，機械学習の一つである強化学習が頻繁に活用される．強化学習は与えられた環境内で最も利益を生むような行動を試行錯誤して学習する手法である．強化学習で主流の一つとなっているのが Q 学習 [1]である．Q 学習は定められた政策に従いつつ，状態と行動の組に対して推定価値である Q 値を与えていく強化学習の手法である．また，Q 学習で連続した空間を扱うようにしたものがファジイ Q 学習 [2]である．しかし，このままでは Q 学習の特徴から，Q 値が最大のものを選び続けてしまうことで局所解が生まれ，学習が進まなくなってしまう問題が発生する．

本研究では，ファジイ Q 学習と UCBQ アルゴリズム [3]を組み合わせた UCB ファジイ Q 学習を提案し，ドットイートゲームに適用する．この手法は，ドットイートゲームにおけるファジイ Q 学習において，未選択の行動を減らし，さらに Q 値が局所解になることを防ぐことで学習を効率的に進めることを目的とする．UCBQ アルゴリズムとは，UCB アルゴリズム [4]という行動選択手法を Q 学習に応用できるように改良した手法である．UCB アルゴリズムは価値推定値の代わりに UCB 値と呼ばれる値が最も高い行動を選択する．またこのアルゴリズムは未選択の行動があれば必ず選択し，価値推定値を獲得することを前提としているため，任意の状態での選択されていない行動が極めて少なくなり，局所解になることを防ぐ．本研究では，ドットイートゲームの一つである，Ms. PacMan に提案手法を適用し，アルゴリズムで用いられる数値を変化させる実験を行うことによって最適な数値を示した．評価は先行研究 [5]の AI とのスコアの比較により行う．実験の結果，提案手法によって未選択の行動が大幅に削減された．

2. 関連研究

今までにもファジイ Q 学習を用いたゲーム AI についての研究が行われてきた．中島ら [2]はファジイ Q 学習によるサッカーエージェントを提案し，連続した状態空間と行動空間を強化学習で扱う際の困難さを決定論的な行動選択手法などで取り扱うことで回避することを可能にした．馬野ら [6]はカーレースゲームにファジイ Q 学習を適用した．このカーレースゲームは 2 次元平面上に置かれた目標を目指しながら得点を競うゲームである．馬場らは「目標までの距離」，「目標 1 との角度」，「カーエージェントの速さ」の三つの連続した属性を状態空間として用い，目標を通過したときのカーエージェントの向きによる追加報酬として向き報酬を与えることでより少ないステップ数で通過すること可能にした．

ドットイートゲームの一つである Ms. PacMan をプレイするゲーム AI の研究も行われている．DeLooze ら [5]はパックマンと敵，ピル，パワーピルとのそれぞれの距離を属性とするファジイ状態空間を用いることで，連続的な状態を離散的にとらえることを可能とし，Q 学習に適用した 2017 年に Microsoft の AI [7]が Ms. PacMan の最高スコアを打ち出したことを発表した．この AI は Hybrid Reward Architecture と呼ばれるアーキテクチャを採用しており，150 以上の単目的のエージェントとそれらから得られる情報から総合的な判断をするトップエージェントから構成される．これを用いることによってより効率的な学習を可能にした．

Supervisor: Prof. Hiroshi Hosobe

3. Ms. PacMan

Ms.PacMan は 1980 年にナムコから発売された PacMan をもとにアメリカで製作されたゲームである(図 1)．このゲームはゴーストと呼ばれる敵に当たらないように餌をたくさん集めてより高いスコアを稼ぐことを目的とする．また，このゲームではパワーピルと呼ばれる，一時的にゴーストを無力化する餌がマップ上に配置されている．ゴーストを無力化することで食べることができ，高いスコアを稼ぐことが可能となる．さらに連続でゴーストを食べることで，200，400，800，1600 点と点数が高くなる．よって，ハイスコアを狙うにはこの点数をうまくとっていくことが鍵となる．

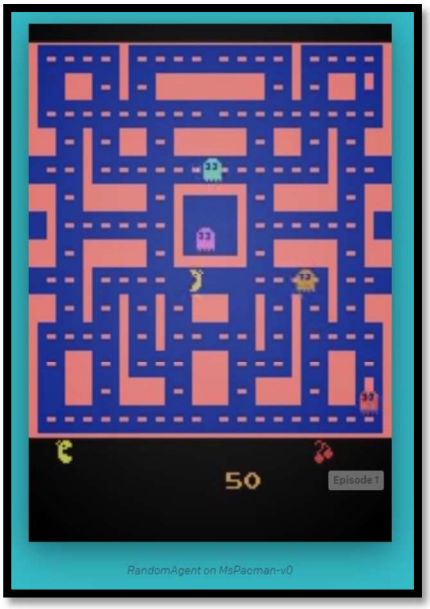

図 1　Ms. PacMan のゲーム画面

PacMan ではゴーストがパックマン(自機)を見つけた際に決定論的に行動するため，AI による攻略方法が確立されていた．それに対して，Ms. PacMan では，ゴーストが一定確率でランダムに行動するため，学習の難易度が上がり，攻略方法が確立されていない．よって IEEE CEC などの国際会議ではこのゲームを題材とする競技が行われていた．

4. 準備

4.1. Q 学習

Q 学習は 1992 年に Watkins ら [1]が機械学習手法の一つとして提案したものである．この手法はエージェント(学習者または意思決定者)がある状態の下でどのような行動をとるべきかという指標である価値推定関数$Q(s,a)$を更新していくことで学習を行う．エージェントは行動した結果に応じて報酬を受け取り，その報酬を用いて$Q(s,a)$を更新する．更新の式は以下の通りである：

$$Q(s,a) = (1-\alpha)\,Q(s,a) + \alpha[R(s,a,s') + \gamma\, maxQ(s',a')]$$

ただし，αは学習率，γは割引率である．$(1-\alpha)$の項は今の Q 値を表し，αの項は学習で用いる値を表している．この式より，報酬が高いほど更新される Q 値が高くなることがわかる．

4.2. ファジイ Q 学習

ファジイ Q 学習は中島ら [2]が 2008 年に提案したものである．Q 学習の状態空間と行動空間をファジイ集合として扱ったものをメンバーシップ関数で表し，メンバーシップ関数から得られるメンバーシップ値を Q 値の算出に用いることで，連続的な状態空間と行動空間を Q 学習で扱うことを可能とした．

DeLooze ら [5]はファジイ Q 学習を Ms. PacMan に適用した．その際，行動空間は離散的であるため，状態空間のみをファジイ集合として扱った．結果としてファジイ集合によってある程度の学習を行うことができたが，この手法だけでは学習時に発生しない状態と行動の組み合わせが多数存在するため，十分な学習ができなかった．Q 学習の特徴により，得られた Q 値の高い行動をひたすら追ってしまうことが学習を妨げた原因と考えられる．

4.3. UCB アルゴリズム

UCBアルゴリズム [4] は価値推定値（Q学習ではQ値）の代わりに UCB 値と呼ばれる値を用いて，その値が最も高いものを選択する．また，ある状態で未選択の行動があった場合，必ず選択し，価値推定値を獲得することが前提とされていることが特徴である．Q 学習における UCB 値は以下の式で表される：

$$UCB(s,a) = Q(s,a) + C\sqrt{\frac{\ln(n)}{N(s,a)}}$$

ただし，nはある状態sでの今までの総プレイ回数，$N(s,a)$はある状態sで行動aを選択した回数，Cは探索の傾向を決める定数である．Cの値が大きいほど積極的に探索を行う．逆に小さいほど，今までの学習の活用を重点的に行うことを意味する．

4.4. UCBQ アルゴリズム

斉藤ら [3]は UCB アルゴリズムを Q 学習に適用した UCBQ アルゴリズムを提案した．アルゴリズムはそのまま Q 学習に組み込んだだけでは問題が生じる．UCB アルゴリズムは未選択の行動を必ず選択するため，次の状態へ遷移不可能な行動を選択し続ける状態に陥る可能性がある．これに対して斉藤らはε-greedy 法を参考にし，一定の確率でランダムに行動を選択させることで問題を解決し，Q 学習に適用できるように改善した．

5. 提案手法

本研究では，行動選択手法の一つである UCB アルゴリズムを Q 学習に適用させた UCBQ アルゴリズムをファジイ Q 学習に組み合わせた UCB ファジイ Q 学習を提案し，Ms. PacMan に適用する．この手法により，未選択の行動を大幅に減らし，学習を効率よく進めることが可能とする．UCB を適用する際，以下を繰り返し行うことで学習を行う．

1. パックマン(自機)と，ゴースト(敵)，ピル(餌)，パワーピル(敵を無力化される餌)とのそれぞれの距離からファジイ状態を得る．
2. εの確率でそのファジイ状態における UCB 値が最大の行動を選択し，実行する．そうでなければランダムに行動を選択する．
3. 距離の数値や行動を行った結果に対する報酬から Q 値の更新を行う．

図 2 は提案手法を疑似コードで表したものである．5~20 行目は 1 回のプレイを学習する際の UCB ファジイ Q 学習のアルゴリズムであり，以下をゲームオーバーになるまで繰り返して学習を行う．6 行目では，εの確率でランダムに，$1-\varepsilon$の確率で UCB により最大の行動を選択する．11 行目では，パックマンとゴースト，ピル，パワーピルのそれぞれの距離の度合いから行動を行った次の状態s'を求める．14 行目では，ファジイ集合からメンバーシップ値を求め，それらを平均した値μ_sを求める．15 行目ではμ_sと報酬Rを用いて Q 値を更新する．17 行目で状態 s の際に選ばれた行動 a に対応するN_aの値を更新し，18 行目でそのN_aの値と更新された Q 値を用いて UCB 値の更新を行う．

```
    procedure UCB fuzzy Q-learning
1   begin
2     initialize Q, UCB, N_a, ε, ∀s ∈ S, ∀a ∈ A,
                numEpisode
3     for cycle := 1 to numEpisode
4       state(s)        #状態の初期値
5       while (not done) do
6         if rand() > ε
7           a ← argmax UCB(s,:)
8         else
9           a ← randomselect
10        end
11        s' ← fuzzyState(pacman, ghost, pill,
                          powerpill)
12        R ← GetReward(s,a)
13        for cycle := 0 to len(Q)
14          μ_s
              ← (membership(pacman,ghost)
              + membership(pacman,pill)
              + membership(pacman,powerpill))/3
15          Q(s,a) ← Q(s,a)
                +μ_s α [R + γ max_{a'∈A(s)} Q(s',a') − Q(s,a)]
16        end
17        N_a ← N_a + 1
18        UCB(s,a) ← Q(s,a) + C √( ln Σ(N_a(s,:)) / N_a(s,a) )
19        s ← s'
20      end
21    end
22  end
```

図 2　UCB ファジイ Q 学習

6. 実装

Ms. PacMan の実装には OpenAIGym を用いた．OpenAIGym は非営利団体である OpenAI が提供している，強化アルゴリズムの開発と評価のためプラットフォームである．OpenAIGym には倒立振子やテレビゲームなど，様々な強化学習の環境が用意されており，その中に ATARI の Ms. PacMan がある．

マップ上のパックマン，ゴースト，ピル，パワーピルの位置情報の把握のために OpenCV を使用した．OpenCV は画像処理，画像解析および機械学習の機能を持つライブラリである．OpenAIGym の Ms. PacMan は色がはっきりしているため，処理を容易に行うことが可能である．それぞれの物体の色を指定し，ピクセルグループを見つけることで，座標を取得している．

7. 実験

パックマン(自機)の行動に対する報酬を調整し，どのような報酬の与え方が効率的な学習に繋がるかを検証する．実験はスコアを比較することで検証・確認を行う．スコアはピル，パワーピル，無力化されたゴーストを食べることで得られるようになっている．学習で用いられる状態sは図 3，図 4 で定義したファジイ集合の組み合わせの計 27 個からなる．行動aは「最も近いピルに向かう」，「最も近いパワーピルに向かう」，「ゴーストから逃げる」の 3 種類から選択され，価値推定関数$Q(s,a)$の要素は全部で$27\times3=81$個あり，学習前の値はすべて 50 としている．

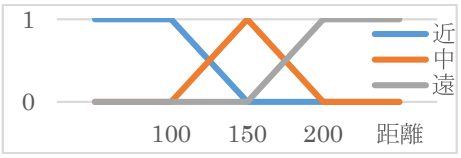

図 3　パックマンとゴーストとの距離のファジイ集合

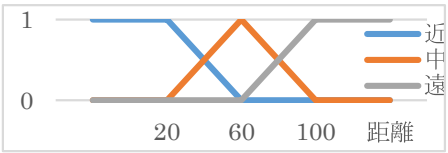

図 4　パックマンとピル(パワーピル)との距離のファジイ集合

実験では報酬R，UCBQ アルゴリズムに使われる定数Cの値を変えることで最適な数値を求める．この時，報酬は行動によって得たスコアを用い，さらにパックマンがゴーストにつかまった際，ペナルティとして-300が与えられる．この数値を定数倍することで Q 値の更新に用いる．

表 1 は実験を行ったパラメータとその結果であるスコアを示している．マシン 0 はC，ランダムに行動する確率を決める定数εを 0 にすることで UCBQ アルゴリズムが組み合わされていない，先行研究と同様にファジイ Q 学習のみのものであり，報酬の与え方のみが異なる．マシン

2〜6 では提案手法である UCB ファジイ Q 学習を適用したものであり，マシン 2，4 では報酬の与え方，マシン 2，5，6 ではCの値が異なる．実験の結果，報酬として$0.03R$を与えたとき，最も高いスコアが得られた．また，Cの値は，できるだけ小さい値の方がスコアをあげることに適していることが分かる．

表 1　実験を行ったマシンのパラメータとスコア

マシン	報酬R	定数C	ε	スコア
0	$0.03R$	0	0	3360
1	$0.03R$	0	0.2	3480
2	$0.03R$	0.01	0.2	3560
3	$0.05R$	0	0.2	3120
4	$0.05R$	0.01	0.2	3300
5	$0.03R$	0.1	0.2	3280
6	$0.03R$	3.0	0.2	2460

マシン 0 では Q 値が初期値である 50 のままとなっている要素がマシン 2，4，5，6 よりも多かった．またマシン 2，4，5，6 ではほとんどの状態と行動の組み合わせで Q 値が初期値である 50 から違う値に更新された．提案手法を用いたマシン同士では特別大きな Q 値の差は見られなかった．

8. 議論

最初に提案手法を用いて探索の度合いを表す定数Cの数値のみを変えたマシン 2，5，6 を比較する．4.3 節で述べた通り，Cの値は大きいほど積極的に探索を行い，小さいほど，今までの学習の活用を重点的に行う．実験より，Cの値が大きくなるほどスコアが低くなっていった．Cの値を大きくしすぎた場合，悪手だとすでに学習されたものでも何度も選択してしまうため，スコアをあげることは難しい．よってCの値を特に大きく設定したマシン 6 は探索に重点をおいた結果，探索と活用のバランスが悪くなったと考えられる．このため，実験で動作させたマシンの中で最もスコアが高かった，Cの値が 0.01 に設定されたマシン 2 が最も学習時の活用と探索を効率よく行ったものであると考えられる．

次に先行研究と同様，ファジイ Q 学習のみを用いたマシンであるマシン 0 と提案手法の中で最もスコアの高かったマシン 2 を比較する．この二つはどちらも報酬を 0.03 倍したものが Q 値の更新で使用されている．よって，今回の実験において Ms. PacMan へ Q 学習を適用する際に適した報酬の値は$0.03R$であることが考えられる．また$0.05R$に設定した際に少しスコアが落ちた原因としては，ピル，パワーピルを食べた際の報酬が多くなったことによりゴーストから逃げることよりもピル，パワーピルを食べることを優先してしまい，結果としてリスクのある行動を多く選んでしまったことが原因であると考えられる．Q 値について比較すると，マシン 2 の Q 値はマシン 1 の Q 値に比べて更新されている要素が多かったことから，マシン 2 は十分な探索を行い，効率的な学習ができたと考えられる．

実験の結果，提案手法の方が 100 点前後上回るスコアを出す結果が得られたものの，提案手法と先行研究との間に大きな差は見られなかった．その原因は，今回のプログラムでは先行研究と同様に 2 面までのマップまでにしか適用されていなかったため，得られるポイントに限りがあったためだと考えられる．よって 3 面以降のマップに適用できるようにすることで提案手法の有用性を改めて確認する必要がある．

本研究で実装したプログラムでは，行動を「最も近いピルに向かう」，「最も近いパワーピルに向かう」，「ゴーストから逃げる」の三つにしていたため，無力化されたゴーストを積極的に食べに行く行動がなかったことでスコアがあまり伸びない結果になったと考えられる．解決策としては，新たに「ゴーストに向かう」などの行動を増やすことが考えられる．

9. おわりに

本論文ではファジイ Q 学習と UCBQ アルゴリズムを組み合わせた UCB ファジイ Q 学習を提案した．実験では Ms. PacMan に適用させ，報酬や UCBQ アルゴリズムの定数C，確率εの数値を変化させることによって最適な数値を示した．

今後の課題は，Ms. PacMan の 3 面以上のマップにプログラムを対応されることで，先行研究の手法とのスコアの差異を明確にすることである．また，行動の選択肢を増やすほど UCB ファジイ Q 学習の行動選択の影響が大きくなると考えられるため，さらに細かく行動を分けることでスコアが上がるかを検証する必要がある．

文　献

[1] C. J. Watkins and P. Dayan, "Q-Learning," *Kluwer Academic Publishers,* 1992.

[2] 中島智晴, 有働晶代, 石渕久生, "ファジイ Q 学習によるサッカーエージェントの行動獲得," 知能と情報, vol. 15, no. 6, pp. 702-707, 2003.

[3] 斉藤晃貴, 野津亮, 本多克宏, "強化学習における UCB 行動選択手法の効果," 第 30 回ファジィシステムシンポジウム講演論文集, pp. 174-179, 2014.

[4] P. Auer, N. Cesa-Bianchi and P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning,* vol. 47, pp. 235-256, 2002.

[5] L. L. DeLooze and W. R. Viner, "Fuzzy Q-Learning in a Nondeterministic Environment: Developing an Intelligent Ms. Pac-Man Agent," *Proc. IEEE Symposium on Computational Intelligence and Games,* pp. 162-169, 2009.

[6] 馬野元秀, 立野宏樹, 伊瀬顕史, "カーレースゲームへのファジィ Q 学習の適用," 第 29 回ファジィシステムシンポジウム講演論文集, pp. 1006-1011, 2013.

[7] H. v. Senjen, M. Fateme, J. Romoff, R. Laroche, T. Barnes and J. Tsang, "Hybrid Reward Architecture for Reinforcement Learning," *arXiv:1706.04208v2,* 2017.