カヌーレーシングのシミュレーション環境における強化学習について

奈良女子大学 理学部 情報科学科 4 回
新出研究室 宮田怜奈

概 要

自律エージェントに関する研究において、連続的な実世界環境では様々な問題に直面するため、それらに対応した行動の自律的獲得が望まれる。このような自律的行動を、環境とエージェントの相互作用を通して学習する手法として、強化学習が適していると考えられる。

先行研究では、連続的な仮想世界におけるエージェントのシミュレーション環境の実現に向け、カヌーレーシングをテストベッドとして強化学習を実装し、自律的に行動決定を行うことが可能になったが、強化学習の方法は固定であり、複数の学習方法を比較することができていなかった。そこで本研究では、複数の強化学習方法を利用可能とする機能の実装を行い、実際に複数の学習方法による比較を行って、カヌーレーシングにおける最適な学習方法の決定を目指した。

1 はじめに

近年、実世界で環境が変化しても、目標達成のために学習や意思決定を行いながら、自律的に振る舞うロボットの実現に向けた研究が進められている。我々は、このような実世界での目標を達成するロボットの実現を目指すため、BDI モデルを用いた自律エージェントの研究を行ってきた。BDI モデルとは、信念 (B)、願望 (D)、意図 (I) と呼ばれる 3 つの心的状態パラメータを用いて自律的な行動決定を行うモデルである。このモデルに基づいた合理的かつ自律的なエージェントを実現したものが BDI エージェントである。

連続的な実世界環境では、離散的な仮想世界とは異なる様々な問題に直面するため、それらに柔軟に対処できる能力が必要となる。しかし、そのための実験をいきなり実世界で行うことは難しい。そのために、BDI エージェントを持つ連続的な仮想世界におけるシミュレーション環境が有用であると考える。先行研究 [4, 2] では、カヌーレーシングをテストベッドとした、連続的な仮想世界におけるエージェントのシミュレーション環境の構築を行い、エージェントが環境を知覚し、その情報をもとに強化学習を行うことで、自律的に行動を選択することが可能となった。しかし、強化学習の方法は ϵ-greedy 法を用いた Q 学習のみに固定されていた。そこで本研究では、より効果的で最適な学習方法を決定することを目指し、[4, 2] に複数の強化学習方法を利用可能とする機能の実装を追加し、カヌーレーシングのシミュレーションの例で実際に複数の学習方法による比較を行った。また、複数の学習パラメータによる比較も行った。

本研究は、奈良女子大学理学部情報科学科 4 回生柚木との共同研究 [3] である。本論文では、カヌーレーシングのシミュレーション環境における最適な強化学習の方法について述べる。

2　強化学習の実装方針

2.1　強化学習

強化学習 [1] とは、『「どの選択肢を選ぶのが正解か」を教えてくれる存在はいないが、実際に試すことによって何らかの報酬が得られる』場合に、「試してみることによって各選択肢の価値を推定する」ことと、「最も推定価値の高い選択肢を得る」ことのバランスを取ることによって、うまい行動を見つけ出そうとするやり方である。すなわち、エージェントが知覚した情報をもとに、どの行為が最も望ましいかを学習することである。

行為の選択には、価値と呼ばれる重みを用い、その価値を更新することで学習は進み、それぞれの学習方式により価値の更新方法は異なる。また、その価値を用いてどのように行動選択をするかでも、学習の性能は異なってくる。

強化学習を行う目的は、獲得する報酬量、つまり通算報酬を最大化することである。そのために、各行為の価値を評価し、その価値を用いて最適な行為を選択する必要がある。

2.2　学習方式

従来は Q 学習のみであった学習方式に、新たに Sarsa を追加した。学習によってより最適な行動をとれるようにするために「ある状態 s においてある行動 a をとったときの、その行動からエピソード終了までの通算報酬」の平均値を推定して保持し、この値を「状態行動価値」と呼び、$Q(s,a)$ と書く。また、行動 a で直接得られた報酬を即時報酬 r とし、状態 s から行動 a を選び、次の状態 s' に遷移したときに選択する状態行動価値を $Q(s',a')$ とする。ここで、α はステップサイズ・パラメータ、γ は割引率である。今回は $\alpha = 0.1$、$\gamma = 0.9$ と設定した。

2.2.1　Q 学習

Q 学習は方策オフ型 TD 制御である。行動の価値の更新は、遷移した状態で最も価値の大きな行動の価値に基づく。Q 学習の更新式は次のようになる。

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\Big[r + \gamma \max_{a'} Q(s',a')\Big]$$

2.2.2　Sarsa

Sarsa は方策オン型 TD 制御である。行動の価値の更新は、遷移した状態で次に選択する行動の価値に基づく。Sarsa の更新式は次のようになる。

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha[r + \gamma Q(s',a')]$$

2.3　行動選択法

従来は ϵ-greedy のみであった行動選択法に、新たにソフトマックス行動選択を追加した。1 つ 1 つのエピソードについて、毎回の学習をそれぞれの行動選択法によって選び実行し、行動のたびに状態行動価値の推定値を更新していくことで学習は行われる。

2.3.1 ϵ-greedy 行動選択

$\epsilon(0 \leq \epsilon \leq 1)$ の確率でランダム選択を行い、$1-\epsilon$ の確率で greedy 選択を行う。

2.3.2 ソフトマックス行動選択

推定価値を等級付けした関数によって行動確率を変化させ、t 回めのプレイにおける行動 a を次の確率で選択する。

$$\frac{\mathrm{e}^{Q_t(a)/\tau}}{\sum_{k=1}^{n} \mathrm{e}^{Q_t(a_k)/\tau}}$$

τ は温度パラメータと呼ばれる。温度が高い場合には、すべての行動が同程度起こるように設定される。また、温度が低い場合には、価値の推定が異なる動作の選択確率の差がより大きく異なるように設定される。$\tau \to 0$ の極限では、ソフトマックス行動選択は greedy 行動選択と一致する。

2.4 実装方法

[4, 2] では、状態とエージェントの行為、および行動の結果の知覚をもとに学習を行うクラスである RiverUpdate を作成していた。本研究では、複数の強化学習方法の比較を行うため、このクラスに、

- ϵ-greedy 法を用いた Q 学習
- ϵ-greedy 法を用いた Sarsa
- ソフトマックス法を用いた Q 学習
- ソフトマックス法を用いた Sarsa

の 4 つの学習方法をそれぞれ実装した。

2.5 環境設定

カヌーレーシングのシミュレーション環境において、エージェントが岸に衝突しないようにカヌーを操り、できるだけ速く川を下りきるような行動を学習させる。途中で岸にぶつかってしまうと失敗となり、エージェントがスタート地点を出発してから、岸にぶつかって止まる、またはゴール地点に到達し川を下りきるまでを 1 つの「エピソード」と呼ぶ。

川の座標は連続値であるが、学習を簡単にするため、学習は図 1 の 1 つのマスを 1 つの状態として、状態ごとに最適な行為を学習する方式となっている。なお、今回は 8×6 のマスに区切っている。

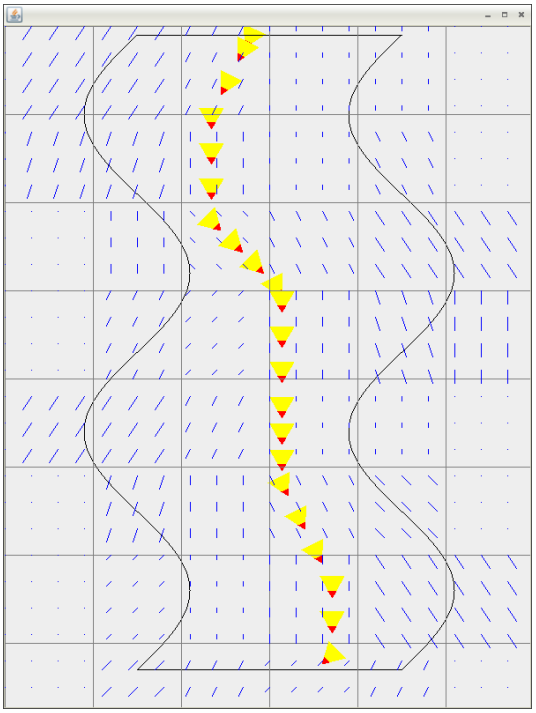

図 1: カヌーレーシングのシミュレータ

2.5.1 報酬値の設定

[4, 2] と同様、報酬値を以下のように設定した。

- 川を下りきったら報酬 100
- 岸に衝突したら報酬 −100
- それ以外なら基本行為 1 回あたり報酬 −1

この報酬値の設定の上で、エピソード終了までの通算報酬を最大化する行動をとることを目指す。

2.5.2 基本行為の設定

[4, 2] と同様、基本行為は 5 つに定めている。

- 何もしない
- 前進
- 後退
- 右へ曲がる
- 左へ曲がる

上記の 5 つの基本行為のうち、どの行為を選択すればよいかをエージェントは学習する。

2.5.3　強化学習の評価

強化学習の方式を比較し評価するにあたり、

- 川から出た回数 (岸にぶつかった回数)
- 平均報酬 (通算報酬の平均)

上記の 2 つの要素を対象とした。試行回数を増やすにつれ、この 2 つの要素の推移を調べ、評価を行った。

2.6　実験手法

今回は各学習方法について 10000 回エピソードを繰り返し、学習を行った。学習中、コンソール画面に「川から出た回数」と「平均報酬」を図 2 のように出力させた。出力は、1, 10, 100, 200, 300, 400, 500, 1000, 5000, 10000 エピソードの終了時点まで行った。

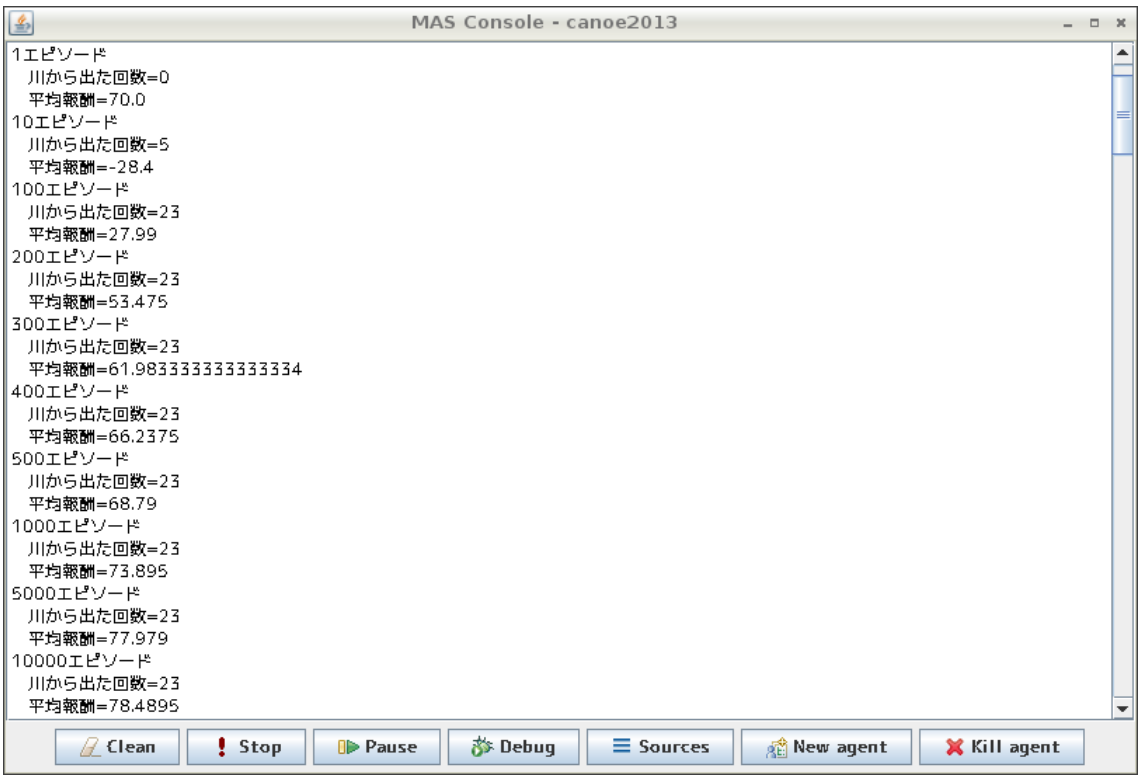

図 2: コンソール画面

図 2 から、「川から出た回数」、「平均報酬」について 10000 回のエピソードをさらに 10 回繰り替えしてその平均のデータをとり、グラフで比較を行った。

3　実験結果

3.1　ϵ-greedy 法のパラメータ

ϵ-greedy 法で ϵ の値を 0.001, 0.01, 0.1 とそれぞれ変えて Q 学習と Sarsa のそれぞれの平均報酬を比較したグラフは次の図 3, 4 のようになった。

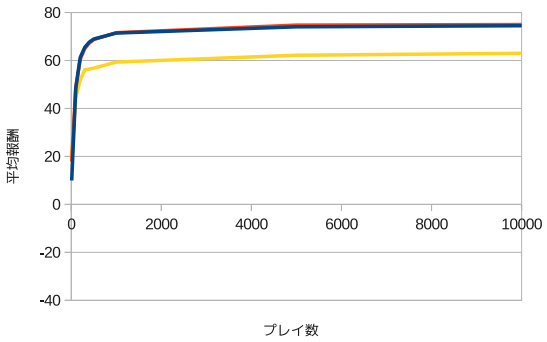

図 3: Q 学習

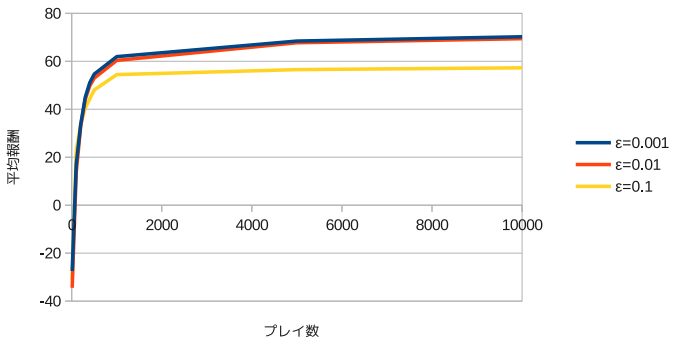

図 4: Sarsa

この結果から、Q 学習と Sarsa のいずれでも $\epsilon = 0.1$ よりも $\epsilon = 0.01$ の方が平均報酬は改善しているが、$\epsilon = 0.01$ よりも小さくした $\epsilon = 0.001$ は $\epsilon = 0.01$ とほとんど類似した。あまりにも ϵ を小さくするとグリーディ以外の行動をほとんどとらないので、特に学習初期において学習効率が下がる。そこで、今回は $\epsilon = 0.01$ に設定し、実験を行った。

3.2　ソフトマックス法のパラメータ

ソフトマックス法で温度パラメータ τ の値を 0.1, 0.5, 1 とそれぞれ変えて Q 学習と Sarsa のそれぞれの平均報酬を比較したグラフは次の図 5, 6 のようになった。

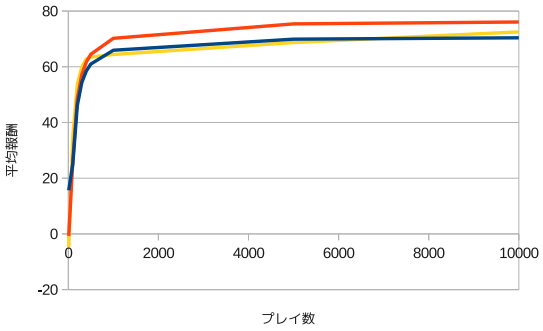

図 5: Q 学習

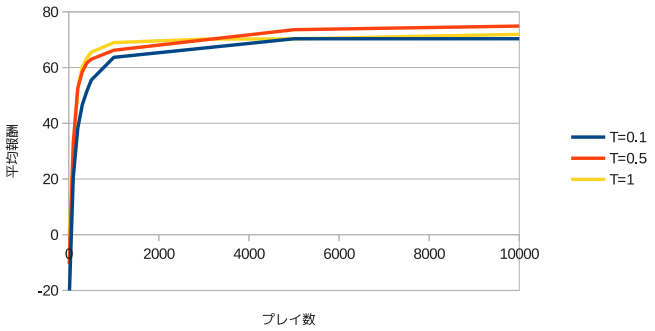

図 6: Sarsa

この結果から、Q 学習でも Sarsa でも 10000 回エピソードを繰り返すと、$\tau = 0.5$ の時が最も平均報酬が高くなった。そのため、今回は温度パラメータを $\tau = 0.5$ に設定し、実験を行った。

3.3　行動選択法の比較

先ほど決めたパラメータの値で行動選択法の ϵ-greedy とソフトマックスの行動選択法の比較を行った。

3.3.1　川から出た回数で比較

Q 学習と Sarsa のそれぞれで川から出た回数を比較したグラフは次の図 7, 8 のようになる。

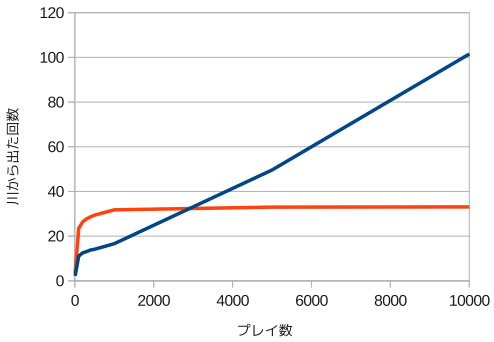

図 7: Q 学習

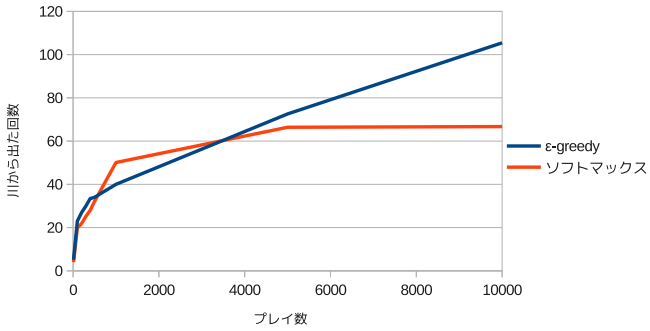

図 8: Sarsa

この結果から、Q 学習でも Sarsa でも、ϵ-greedy はプレイ数が増えるにつれ、川から出た回数も比例して多くなったが、ソフトマックスは途中からほとんど川から出た回数が増えなくなっている。

これは、ϵ-greedy の特徴として、パラメータ ϵ に与えた確率だけランダム選択を行ってしまうため、エピソードを繰り返すと川から出る回数は増え続けてしまうと考えられる。対して、ソフトマックスの特徴として、最も価値の高い行動に最も高い選択確率が与えられるため、プレイ数が増えると価値の高い行動をとりつづけ、川から出る回数が増えなくなったと考えられる。

3.3.2 平均報酬で比較

Q 学習と Sarsa のそれぞれで平均報酬を比較したグラフは次の図 9, 10 のようになる。

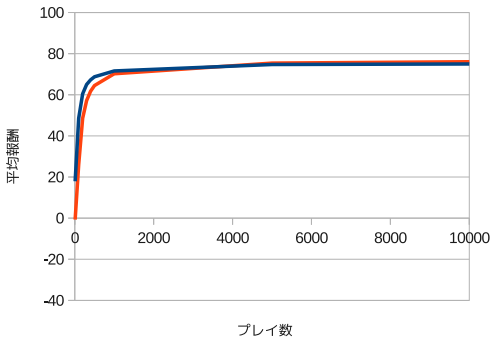

図 9: Q 学習

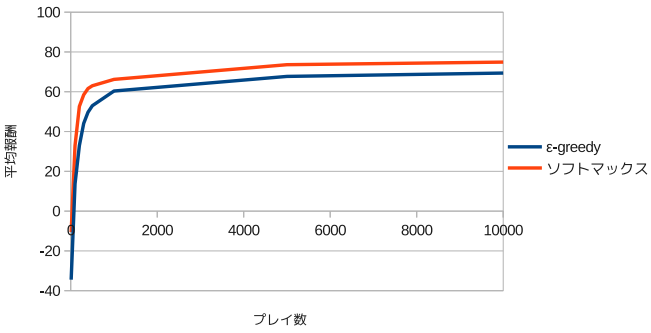

図 10: Sarsa

この結果から、Q 学習の方では ϵ-greedy とソフトマックスに大差は見られなかった。しかし、Sarsa の方ではわずかにソフトマックスの方が平均報酬が上回った。

3.3.3 行動選択法の評価

ϵ-greedy とソフトマックスの比較を行った結果、Sarsa で比較した場合ソフトマックスの方が平均報酬が高くなったこと、また、カヌーレーシングのシミュレーションにおいて、川から出てしまう (岸にぶつかってしまう) という結果は望ましくないこと、2 つの点をふまえて、ϵ-greedy よりもソフトマックス行動選択が望ましいという評価を下した。

3.4 学習方式の比較と評価

ソフトマックス行動選択を用い、Q 学習と Sarsa の 2 種類の学習方式の比較を行った。

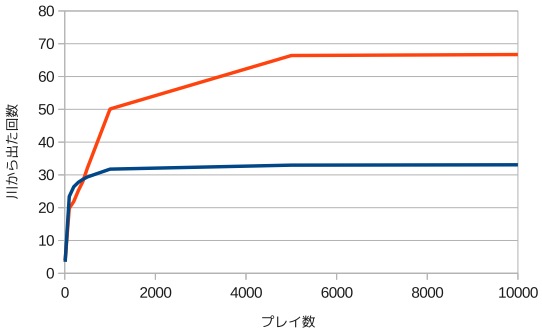

図 11: 川から出た回数

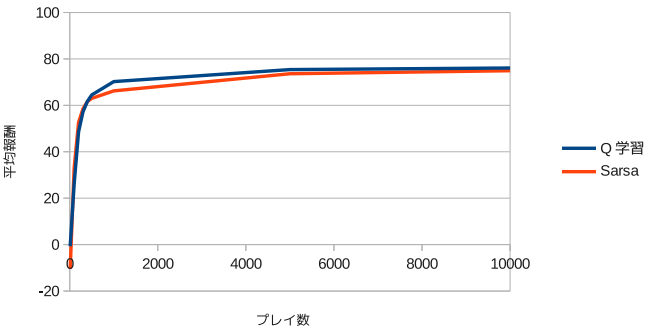

図 12: 平均報酬

図 11, 12 のグラフから、平均報酬の推移は類似したが、川から出た回数は Q 学習の方が少ないという結果となった。そのため、今回の実験結果として、Q 学習の方が望ましいという評価を下した。

4　まとめ

本研究により、連続的な世界におけるエージェントのシミュレーションにおいて、複数の強化学習の比較が可能となった。これによりカヌーレーシングの例では、行動選択法にはソフトマックスを用い、学習方式には Q 学習を取り入れる強化学習が望ましいといった結果を得られた。

現段階での実装は Q 学習と Sarsa の 2 種類のみの学習方式に留まっており、さらに多くの学習方式の実装を目指すことや、パラメータの値を変えた実験をより多く行うこと、カヌーレーシング以外でのシミュレーションで強化学習を行い、シミュレーションごとに最適な強化学習の方法を確立させることなどが今後の課題として挙げられる。

5　謝辞

本論文の執筆及び研究を遂行するにあたり、いつも親身にご指導下さった新出尚之准教授、また亀村美佳先輩に深く感謝し、厚く御礼申し上げます。また新出研究室の皆様に感謝の意を表します。ありがとうございました。

参考文献

[1] Richard S.Sutton and Andrew G.Barto. *Reinforcement Learning.* A Bradford Book, 1998. (三上 貞芳 and 皆川雅章 共訳, 「強化学習」, 森北出版, 2000).

[2] 林果穂. 連続的なシミュレーション環境でのエージェントの学習と意思決定の実装につ いて. 2013 年度卒業論文, 奈良女子大学理学部情報科学科, 2014.

[3] 柚木静香. 実世界のエージェント構築のためのシミュレーション環境の実現. 2014 年度卒業論文, 奈良女子大学理学部情報科学科, 2015.

[4] 亀村美佳. エージェントの知覚と学習による行動決定の実現. 2013 年度卒業論文, 奈良女子大学理学部情報科学科, 2014.