ISSN 2186-5647

P-15

自己組織化マップを用いた教示データの自動生成による強化学習の高速化手法の提案

日大生産工　○檜山昌弘　　日大生産工　山内 ゆかり

1 まえがき

強化学習は、エージェントが未知の環境内で試行錯誤を繰り返すことにより環境に適応した制御法を学習する手法だが、状態数が多い環境では膨大な学習時間が必要になる。

収束を早めるための学習の補助として、館山らの教示データを用いた自己組織化マップを利用した強化学習[1]があり、学習の高速化においては有効性が確認されている。人により用意された教示の導入は教師なし学習である強化学習の本質を損ねている恐れがある。

そこで本研究では，教示データを事前に用意することなく、エージェントが自らの行動経歴より自動生成し、学習の高速化を図ることを目指す。

2　自己組織化マップを用いた教示データによる強化学習

自己組織化マップを用いた教示データによる強化学習とは、マップを考慮した教示データを入力とし、自己組織化マップで学習する。その学習で得られた幅広い範囲の状態空間への教示結果を強化学習の初期値として、強化学習を行うというものである。概念図を図１に示す。

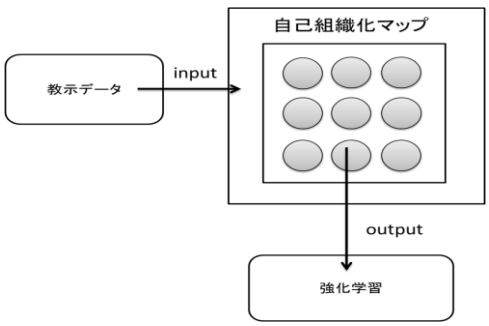

図 1.自己組織化マップを用いた強化学習

3　実験環境及び提案手法

マップは縦12、横18の範囲で図2に示した通りスタートを(0、6)、ゴールは(18、12)に配置した。エージェントは上下左右の4方向の行動ができ、行動選択はルーレット選択を用いる。ゴールしたら報酬を受け取り、Q学習を用いて学習し、スタート位置に戻る。マップにはエージェントの動きを遮る障害物が3つ用意されている。

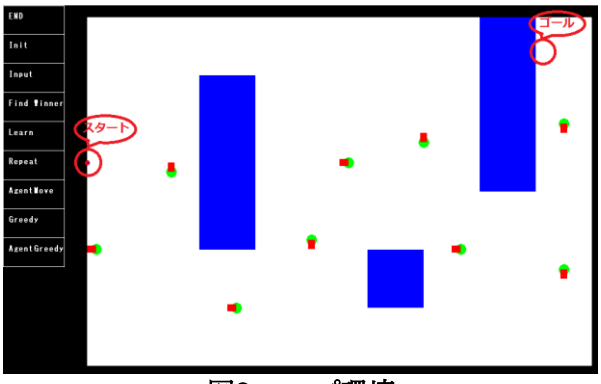

図2.マップ環境

3．1　従来手法

教示データは、マップ上で行動できる座標に与えられており、行動する方向を矢印で表している。従来手法の教示データは図2に示されたように、手動で9個与えられている。

3．2　提案手法

本研究では、強化学習の本質である教師なし学習を行うために、自らの行動履歴を利用した教示データの自動化を目指す。自動生成のアルゴリズムを下記に示す。

1. 事前知識を持たないエージェントが試行錯誤によりゴールするまでの行動を記録し、ゴール後、報酬をもらい、スタート位置に戻る。
2. 行動履歴より迂回経路を削除し、最短経路を作成する。
3. 最短経路における同一行動の連続を調査する。

具体的には、5回中4回同一の行動が連続した場合に、教示データとしてその行動の方向及び中心の座標を記録する。

本実験で用いた各係数を示す。

自己組織化マップ

ニューロン数　2500(50×50)
近傍初期値　40
学習係数　0.4
学習回数　5000

強化学習

報酬　10
割引率　0.99
学習係数　0.5
行動限界数　20000

Automatic Generation of Teaching Signals for Reinforcement Learning Using Self-Organizing Map

Masahiro HIYAMA and Yukari YAMAUCHI

初回ゴール後に自動生成された教示データを図3に示す。教示データ生成数は少ないが、行動履歴より従来手法の手動で教示データと類似した教示データを自動生成することができた。

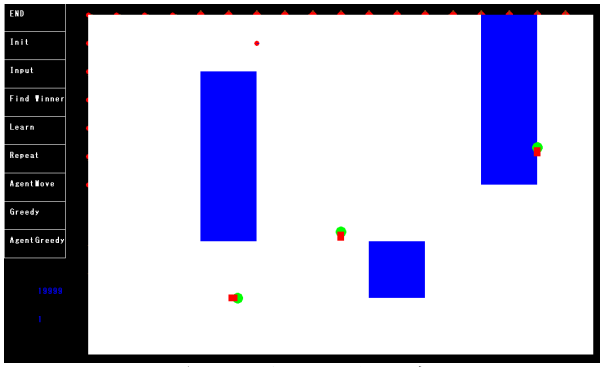

図3.自動生成した教示データ

4　実験結果

本研究では、通常の強化学習(Q学習)と教示データを手動で作成する従来手法と教示データを自動生成する提案手法の3つの環境で実験を行った。10試行までの最小行動数及び総行動数、100試行までの最少行動数及び総行動数、さらに最短経路が得られるまでの試行回数の3つで比較した。それぞれの実験を10回行った平均を表1、2、3に示す。

表１：10試行までの最小行動数及び総行動数

	10試行までの最小行動数	10試行までの総行動数
通常強化学習	313.6	33761.8
従来手法	67.6	2106.4
提案手法	110.2	40744.4

表1より、教示データを用いて自己組織化マップでQ値に偏りを持たせることによって、平均行動数及び総行動数は通常強化学習よりも少ないことがわかり、教示データの有効性がわかる。提案手法は従来手法よりも行動数は多いが、通常強化学習よりは、行動数が少ないことがわかり、自動生成による教示データの有効性が確認できた。提案手法の総行動数が多いのは、初期において行動限界数に達し、教示データがなかなか作れないことが原因と考える。

表２：100試行までの最小行動数及び総行動数

	100試行までの最小行動数	100試行までの総行動数
通常強化学習	115.8	196865.7
従来手法	43.2	14770.6
提案手法	48.8	62621.4

表2より、提案手法の総行動数が通常強化学習の総行動数よりも少なくなり、教示データの有効性が確認できた。

表3.：最短経路が得られるまでの試行回数

	最短経路が得られるまでの試行回数
通常強化学習	48.5
従来手法	566.7
提案手法	511.6

表3より、通常強化学習よりも従来手法及び提案手法のほうが最短を導き出すまでに多くの試行を要することがわかる。このことから、最短を導き出すことにおいては、教示データが妨げになっていることがわかる。

5　まとめ

本研究では、強化学習(Q学習)におけるQ値を自己組織化マップで偏りを持たすために手動で教示データを作成した従来手法を発展させた、教示データを自動生成する研究を行った。学習初期では、教示データの有効性が確認できたが、学習が収束に近づくにつれ、妨げになっていることがわかった。原因としては自己組織化マップを用いてQ値に偏りを持たせて通常の強化学習を行うため、学習する際に、学習によってQ値が更新されて、自己組織化マップでのQ値の偏りが壊れてしまうためであると考えられる。

今後の課題として、自動生成する教示データを従来手法のような教示データの生成を目指し、学習初期での有効性を学習後半にも及ぶようにしたい。

「参考文献」

[1]館山武史、 川田誠一、 小口俊樹、 自己組織化マップを用いた教示による強化学習の高速化手法の提案、 日本機械学会論文集(C編)70巻694号、 (2004-6)

[2]Richard S.Sutton、Andrew G.Barto(著)、三上貞芳、皆川雅章(訳)、 強化学習、 森北出版(株)、(2000/12)