2K-06

自律ロボットの実時間強化学習

Real time reinforcement learning for autonomous robots

太田　悠介　　ハルトノ　ピトヨ

Yusuke Ota　　Pitoyo　Hartono

中京大学　工学部　電気電子工学科

Department of Electrical and Electronic Engineering, School of Engineering, Chukyo University

1. はじめに

近年，人工知能がブームとなり，将来人間とロボットが共存する世界が予想できる．人間が暮らす環境は常に変わり続けるため，前もってロボットの行動をプログラミングすることは困難である．そこで，ロボットに学習能力を搭載することで，動的な環境においても，ロボットが自ら動作戦略を獲得することができる．本研究では，機械学習法をロボットに適用し，実環境下での実時間学習を試みる．

機械学習には，理想的な出力を提示する教師付き学習と理想的な出力を提示しない教師なし学習，結果のみを評価する強化学習に大きく分けられる．動的な環境において，教師信号を常に必要とする教師付き学習の実現は困難なため，ここでは強化学習を搭載することを考える．強化学習によるエージェントの行動獲得は，できるだけ少ない試行回数で完了することが望まれる[1]．しかし，強化学習の代表的なアルゴリズムである Q 学習では，価値関数テーブルを作る事に計算資源を多く必要とするため，計算と電力資源が限られた小型ロボットの実時間学習には適していない．そのため，多くの計算資源を必要としない独自の学習アルゴリズムの開発を行う．ここでは，強化学習で訓練できる階層型ニューラルネットワークの構築を行う．学習方法は[2]に基づいて構築し，その汎化性を評価するために様々な評価関数に関して実験を行い，その成果を本論文で報告する．

2. 実時間強化学習

本研究で用いるリカレントニューラルネットワーク(RNN)を図 1 に示す．提案する RNN はロボットのセンサ情報とフィードバックされた過去の出力値を入力として行動を生成する．

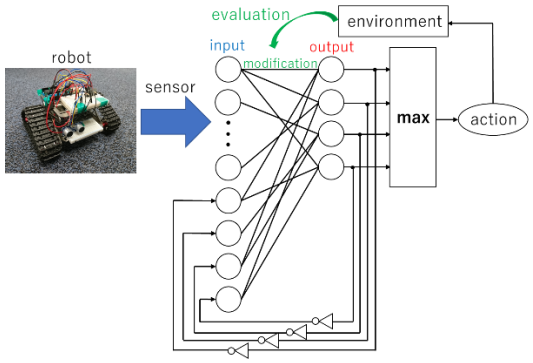

図 1：ニューラルネットワークの概要

時刻 t における i 番目の入力ニューロンの値を $x_i(t)$，j 番目の出力ニューロンの値を $y_j(t)$，i 番目の入力ニューロンと j 番目の出力ニューロン間の重みを $w_{ij}(t)$，入力ニューロン数を n とする．また，j 番目の出力ニューロンのポテンシャルを $I_j(t)$，閾値を $\theta_j(t)$ とする．出力値 $y_j(t)$ は以下の式で計算する．

$$I_j(t) = \sum_{i=1}^{n} w_{ij}(t)x_i(t) - \theta_j(t) \quad (1)$$

$$y_j(t) = \frac{1}{1+e^{-0.001I_j(t)}} \quad (2)$$

その後，式(3)により，ロボットによって行動が実行される．

$$\mathrm{act}(t) = action\left(arg\max_k y_k(t)\right) \quad (3)$$

$action(i)$ は i 番目の出力ニューロンに対応する行動を実行する関数である．ロボットがとった行動が良い行動だったのか，悪い行動だったのかを，評価関数で決定し，評価関数が増加するように重みの更新をする．

・ロボットがとった行動が良い場合，重みの更新を以下に示す．

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.

$$w_{ij}(t+1)=\begin{cases} w_{ij}(t)+\eta & if\ j=arg\max_k y_k(t) \\ w_{ij}(t)-\eta & otherwise \end{cases} \quad (4)$$

式(4)はロボットが良い行動を実行できたため，類似した入力に対し，選ばれた行動を強化すると同時に他の行動を抑制する意味を持つ．

・ロボットがとった行動が悪い場合，重みの更新を以下に示す．

$$w_{ij}(t+1)=\begin{cases} w_{ij}(t)-\eta & if\ j=arg\max_k y_k(t) \\ w_{ij}(t)+\eta & otherwise \end{cases} \quad (5)$$

式(5)ではロボットの行動が悪いため，類似した入力に対し，その行動が選ばれないように，それに対応する出力値を減らすと同時に，他の行動に対応する出力ニューロンの値を増加させるように重みの更新を行う．更新式(4)(5)は正の入力値を仮定し，ηは正の定数である．

ここで提案する学習法は，入力に対し常に取りうる行動が競合し，高い評価値が期待できる行動が実行される．また，学習過程において，とるべき行動を提示せず，その評価のみを与えることで強化学習を実行することができる．

3. 実験

図 2 に示す両端に障害物のある直線的な環境で実験を行った．

図 2：実験環境

実験では，障害物回避をタスクにフィードバックを有しないニューラルネットワーク(NN)と RNN の比較を行った．その学習過程を図 3 に示す．

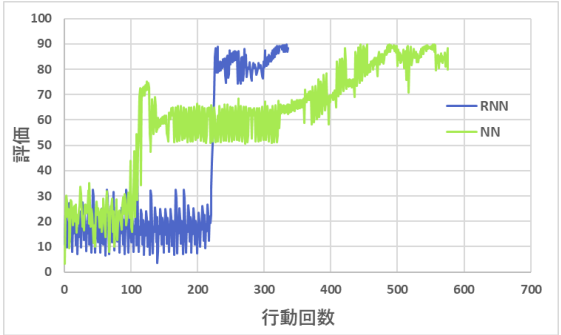

図 3：障害物回避実験

学習回数に従い，評価値が大きくなっていることが分かる．また RNN を使うことで，学習が速くすすむことも分かる．

汎化能力を評価するため，次の実験ではロボットが連続して同一の動作を実行しながら，障害物回避をする学習を行った．ここで，用いる評価関数は式(6)に示す．

$$L(t)=\lambda\times\text{連続性}+(1-\lambda)\times\text{安全性} \quad (6)$$

この式の”連続性”とはロボットが現在の行動と一つ前の行動が同じならば高評価を，”安全性”とはロボットが障害物から遠いところで行動したら高評価を与えるというものである．λは経験的に決定する正の定数($0 \le \lambda \le 1$)．図 4 に学習過程(RNN)を示す．

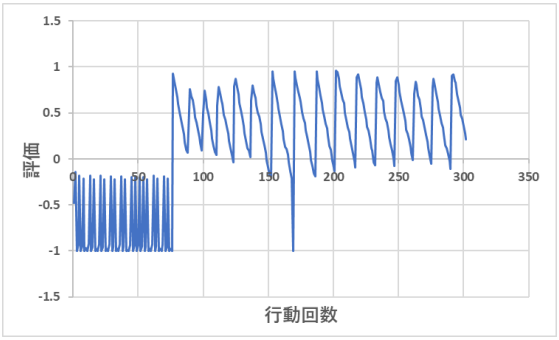

図 4：評価関数

図 4 より，学習時間が進むにつれ，ロボットが障害物に近いところで引き返した時には大きい値をとり，障害物に近づくにつれて，値が小さくなっていく動作が周期的に行われていることが分かる．

4. まとめ

本研究において，ロボットの実時間強化学習に用いることが可能な階層型ニューラルネットワークの開発を行った．実験では，いくつかの評価関数を用いてその汎化性の評価を行った．今後，動的な環境下での実験を行い，計画性を必要とする，より複雑なタスクに対応できるために学習アルゴリズムの改良を行う．

参考文献

[1]小谷直樹,布引雅之,谷口研二,”強化学習における状態数を抑制するクラスタリング方法”,システム制御情報学会論文誌,Vol.22,No.1,pp.21-28(2009).

[2]P.Hartono and S.Kakita,”Fast reinforcement learning for simple physical robots”,Memetic Computing 1(4),pp.305-313(2009).

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.