

Q-Learning を用いた連続空間による RBF の検討

○正 竹原 直美 (函館高専)、正 石若 裕子 (函館高専)

要旨

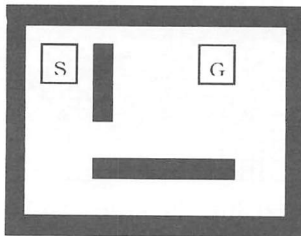
本研究は、強化学習を用いて連続空間内での経路探索を行う事を目的とする。今回は連続空間へ Q 学習を適用するために RBF を用い、シミュレーション結果から Q 値の広がりに対する学習効率の検討を行う。また、GA を用いることにより最適な σ の値を求める実験を行った。

1.はじめに

本研究の目的は、強化学習法の一手法である Q 学習^[1]によって連続空間の経路探索を行うことである。ここで Q 値の更新を連続的に行うため RBF^[2]を用いる。この時 σ の値によって学習効率に変化するため (値によっては収束が得られないことが示されている^[1]) 検討が必要となる。さらに GA^[3]を用いて最適な σ の値を得るために行った実験の結果を示す。

2.問題空間の記述

問題空間は 2 次元連続空間とする (図 1)。エージェントは円形で現在の座標 (x, y) ($s \in S$) を状態として持ちと半径 r の情報を持つ。また、移動方向は 8 方向 (図 2) で 1 ステップ当たりの移動距離とエージェントの半径は固定とする。エージェントは行動選択と Q 値の更新において移動可能な範囲までの状態を観察するものとする。



□ : START
□ : GOAL

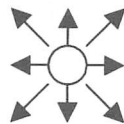


図 1. 2 次元連続空間 図 2. エージェントの移動方向

3.手法

3.1 Q 学習

Watkins^[4]の提案による強化学習法で、状態値と行動を対にして考え、その評価を見積もる。これを Q 値と呼ぶ。Q 学習では状態値と行動を対にして次のステップにおける行動の最大値を基に Q 値を更新していく。更新式は以下ようになる。

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

s : 現在の状態値 s' : 次の状態値

a : 現在の行動 a' : 次の行動
 α : 学習率 γ : 割引率
 r : 環境からの報酬
 \max : 行動集合から、その状態に対する最大の Q 値を獲得する関数。

エージェントは始めスタート位置にあり、1 ステップごとに周囲の状態を観測し移動可能であれば決められた距離だけ移動する。

行動選択には ϵ -グリーディー方策を用いた。これは ϵ の確率でランダム行動をとり、それ以外では現在の状態から最も Q 値が高い方向への行動を選択する (グリーディー探索) ものである。これにより局所性最適解に陥るのを防ぐ。

ゴール到達または指定されたステップ数により 1 エピソードが終了する。Q 値の更新時に用いられる $Q(s, a)$ はエージェントと重なる部分の Q 値の平均値とする。

連続的に Q 値の更新を行うために、RBF を用いた。RBF (Radial Basis Functions) とはガウス分布に代表される関数族である。RBF 法は基本的にはデータ点 $\{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} \mid i=1, \dots, N\}$ が与えられたときに、これを補間する関数 F として、条件

$$F(x_i) = y_i (i=1, \dots, N)$$

を満たすものを、入力データ点 $\{x_i\}$ に “中心” を持つような RBF をデータ個数分だけ配置した

$$F(x) = \sum_{i=1}^N c_i h(\|x - x_i\|)$$

という形式で得ようとするアプローチである。ここで h は RBF である。

今回は RBF として知られているものの中からガウス分布を用いた。

$$h(r) = e^{-\frac{r^2}{\sigma^2}}$$

3.2 遺伝的アルゴリズム (GA : Genetic Algorithm)

GA は John Holland により 1975 年に発表された最適化手法の一つで、生物の進化と遺伝のメカニズムを工学的にモデル化したものである。GA には交叉・突然変異・淘汰といった基本的な操作がある。

GA における選択方法にはいくつかの手法があるが、ここでは適合度の高さから大きい順にランキングを決

めその順位に応じて保存される個体を決定させるランキンング選択という手法を用いた。本実験において、遺伝子は σ の値であり個体数は20、世代数は30、交叉方法は一点交叉である。また σ の値(0.1~6.4で変化量は0.1)を6ビットの遺伝子で表現した。

用意された個体に対する適合度は、その個体(遺伝子)の値でQ学習を行うことによって得られる。今回は指定されたエピソード数におけるゴール数によって適合度を決定した。

4.実験

今回 $\sigma=1.6, 4.0$ の場合によるQ学習と、GAによって σ を求めるという2つの場合について実験を行った。この時の環境は 250×300 の2次元連続空間で、各パラメータは学習率 $\alpha=0.4$ 、割引率 $\gamma=0.2$ 、 $\epsilon=0.1$ 、エージェントの半径 $r=5$ 、移動距離4とした。また最大ステップ数は5000回、最大試行回数は1000回で、更新にはRBFを用いてゴール時の報酬は最大(中心) $r=1$ となるように与えた。

σ の値による学習曲線を見ると、GAによって求めた $\sigma=3.1$ の場合が最も早く収束しているのがわかる。これによりGAを取り入れることにより最適の σ の値が得られたといえる。また、図4、5、6はともにゴール数が50回の時の経路を示したものである。これからも $\sigma=3.1$ の場合が最も早く経路が確定できていることが確認できる。

5.まとめ

今回は σ をパラメータとして与えた場合と、GAによって求めるという2つの場合についての実験を行った。どちらも同様の結果が得られ、GAでの結果が妥当であることが確認出来た。また今回の実験によって、連続空間において経路探索をするさいにRBFの σ の値によって学習効率が向上することがわかった。今後さらに大規模な空間や実ロボットへの適応を考慮するにあたって、他の空間においても同様の実験を行い空間の大きさや障害物などにより σ の値をどのように設定すればよいか検討する必要がある。

6.参考文献

- [1] R.S.Sutton and A.G.Barto : Reinforcement Learning, The MIT Press, pp 135-150, 1998
- [2] 平野 広美: 応用事例でわかる遺伝的アルゴリズムプログラミング
- [3] 丸山 稔: Radial Basic Functions を用いた学習ネットワーク-ニューロコンピューティングに対する新しいアプローチ, システム/制御/情報, Vol.36, No.5, pp.332~329, 1992
- [4] C.H.J.C, Watkins : Q-Learning, Machine Learning 8, pp 279-292, 1992

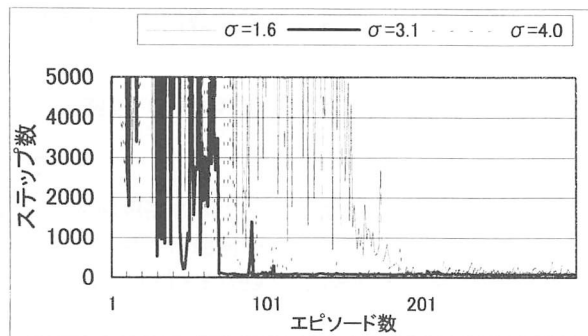


図3 学習曲線の比較 ($\sigma=3.1$ はGAによる結果)

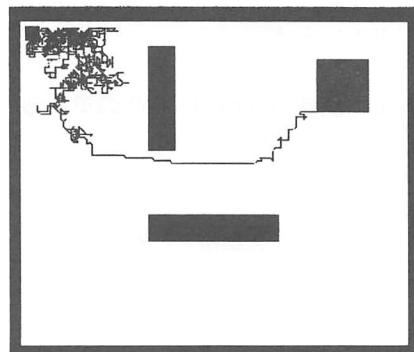


図4 $\sigma=1.6$ の時の経路 (50ゴール目)

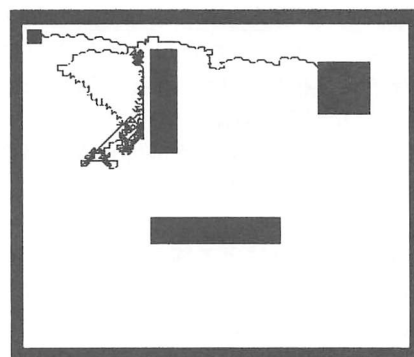


図5 $\sigma=4.0$ の時の経路 (50ゴール目)

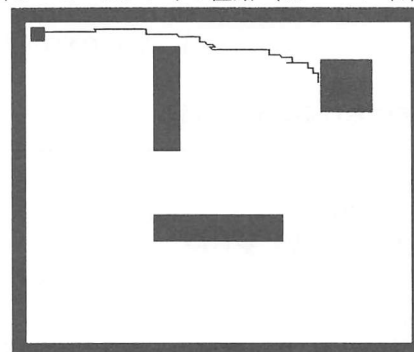


図6 $\sigma=3.1$ の時の経路 (50ゴール目)