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

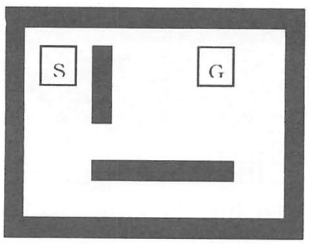
本研究は，強化学習を用いて連続空間内での経路探索を行う事を目的とする．今回は連続空間へQ学習を適用するためにRBFを用い，シミュレーション結果からQ値の広がりに対する学習効率の検討を行う．また，GAを用いることにより最適なσの値を求める実験を行った．

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[3]を用いる．この時σの値によって学習効率が変化するため（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[2]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∈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。



S：START
G：GOAL

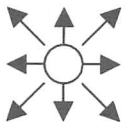


図1. 2次元連続空間　　図2. エージェントの移動方向

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。更新式は以下のようになる。

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

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$\{(x_i,y_i)\in R^d\times R \mid i=1,\cdots,N\}$が与えられたときに、これを補間する関数Fとして、条件

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

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．

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

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σ=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α=0.4，割引率γ=0.2，ε=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σ=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σ=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

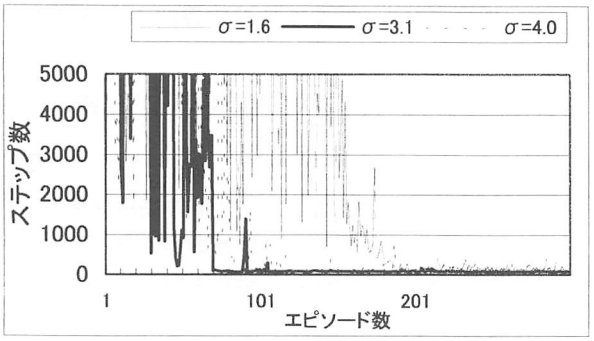


図 3　学習曲線の比較（σ=3.1 は GA による結果）

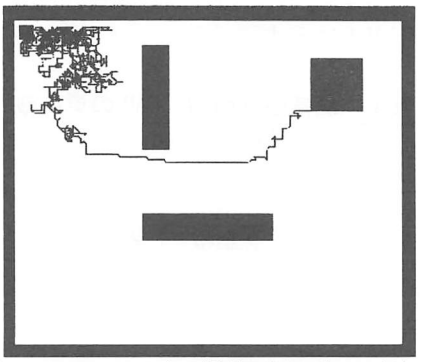


図 4　σ=1.6 の時の経路（50 ゴール目）

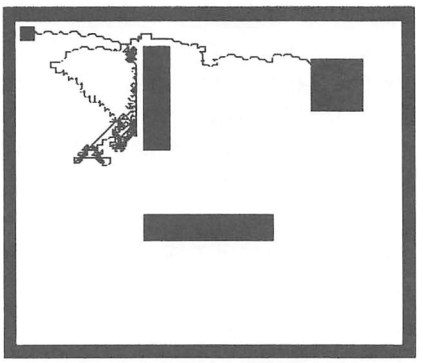


図 5　σ=4.0 の時の経路（50 ゴール目）

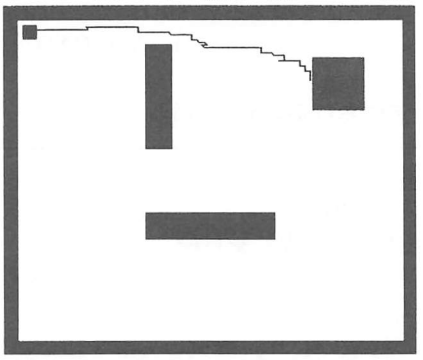


図 6　σ=3.1 の時の経路（50 ゴール目）