強化学習を用いたエージェントの迷路探索問題

○正　竹原　直美（函館高専）、正　石若　裕子（函館高専）

要旨

本研究の目的は、エージェントの迷路探索問題をQ（λ）を用いて行い、λの値に対する結果の比較により、問題に対してλの与える影響力などについて考察することである。Q（λ）とは、強化学習の基本的なメカニズムの1つである適格度トレースとQ学習とを組み合わせたもので、Q学習に比べさらに効率的に学習する事のできそうな一般的手法を得ることができる。今回は17×17、27×27の迷路に対し、λ＝0、5、10の場合について実験を行った。

1．はじめに

本研究の目的は、ロボット制御のアプローチ方法の一つである強化学習、その中でも特にQ（λ）を取り上げ、エージェントの迷路探索問題を様々なλで行い、問題空間の大きさによって結果にどのような変化があらわれるか、さらに実験結果について、将来的に実ロボットに適用することを視野に入れた上で考察を行う事である。今回は2種類の迷路（17×17、27×27）に対して、λをそれぞれ0,5,10と変化させた場合についての実験・考察を行った。

2．問題空間の記述

問題空間は2次元格子状の迷路である。各セルは4つの状態をもち、エージェントは障害物には進入できないものとする（図1）。今回は17×17、27×27の2つの迷路について実験を行った。

エージェントは始めスタートポイントに在り、1ステップにおいて1セルだけ移動しながらゴールポイントを目指す。エージェントの移動方向は上、下、左、右の4方向とする（図2）。

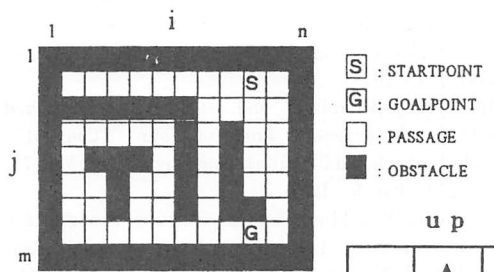

図1 迷路

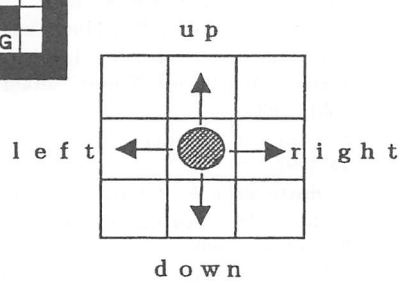

図2 エージェントの移動方向

3．Q－学習とQ（λ）

3．1　Q－学習（行為一価値関数）

Watkins［1992］の提案による強化学習法で、状態値と行動を組にして考え、その評価を見積もる。これをQ値と呼ぶ。Q学習では、方策はもたずに、状態値と行動を組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。

更新方程式は以下のようになる。

$$Q(x,a) \leftarrow Q(x,a) + \alpha\left[r_{t+1} + \gamma \max Q(y,a') - Q(x,a)\right]$$

x：現在の状態値　　y：次の状態値

a：現在の行動　　a'：次の行動

α：学習率　　γ：減衰率

r：環境からの直接報酬

max：行動集合から、その状態に対するQ値を獲得する関数。

行動の選択には、Q値をボルツマン分布とし、確率的に行動を選択する方法を用いる。行動確立Pの式を以下に示す。

$$P(x,a) = \frac{\exp(Q(x,a)/T)}{\sum_{a' \in A} \exp(Q(x,a')/T)}$$

A：行動集合　　T：温度係数

ここで温度係数Tの値が高くなるとQ値の差が反映されにくくなる。

3．2　Q（λ）

強化学習の基本的なメカニズムの1つである適格度トレース（事象に関する記憶パラメータが学習上の変化に対して適格であるとの記録をのこす）とQ学習とを組み合わせたもので、Q学習に比べさらに効率的に学習する事のできそうな一般的手法を得ることができる。

4．方法論

迷路問題では次の①から④のステップを繰り返す。
①迷路の状態を認識する。
（スタート、ゴール、障害物など）
②強化信号の重みを更新する。
（更新式よりルックアップテーブルを生成）
③エージェントが行動を起こす。
④報酬を与える。

5．実験

17×17、27×27 の迷路に対し、λ =0、5、10 の場合について実験を行い比較する。今回、学習率・報酬の割引率は

α =0.9、 γ =0.8

に固定し、最大ステップ数は 5000 回、最大試行回数は 1000 回とした。また全ての λ においてゴール時の直接報酬値 r は 1.0 とし、λ =5,10 の時については、1.0～0 まで値を均一に減少さていく方法を取った。

17×17 の迷路での結果を図５に、27×27 の迷路での結果を図６に学習曲線として示す。

17×17 の迷路においては、λ の値が 0,5,10 のどの値に対しても試行回数が５０回程度までの早い段階でステップ数が収束するという結果が得られた。

それに対して 27×27 の迷路では、λ =0 の場合試行回数が 1000 回程度、またはそれ以上にならなければステップ数が収束しない、つまり学習効果が得られないという結果になった。しかし、λ =5,10 については比較的大きな変化はなかった。

6．まとめ

今回のように、迷路や λ の値の小規模な変化においても、学習効果の違いは大きく現れた。これにより、さらに大規模な迷路や、実ロボットへの適用を考慮した場合、λ の値が問題解決へ大きな影響を与えると思われる。今後、エージェントの取り出し問題、さらに実ロボットへの適用を実現するには、より大きな問題空間と λ の値において実験を重ねる必要性があると考えられる。

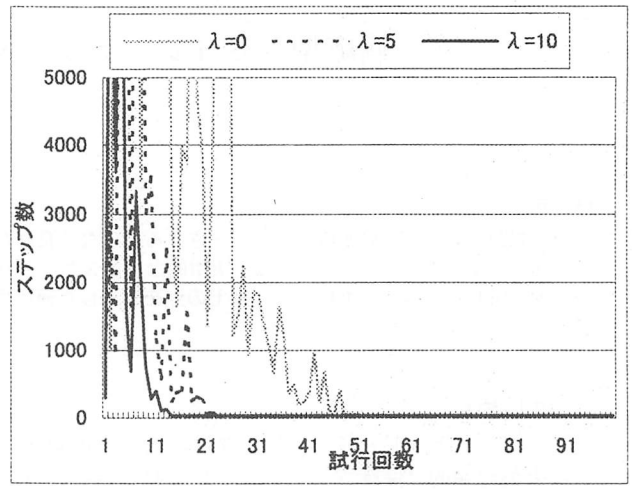

図５　学習曲線の比較（17×17）

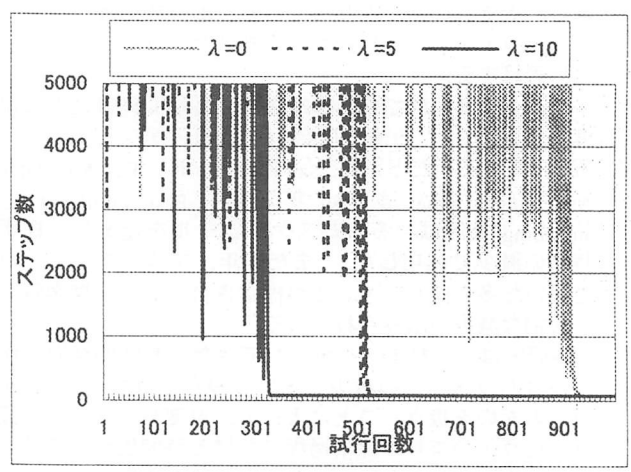

図６　学習曲線の比較（27×27）

7．参考文献

[1] R.S.Sutton and A.G.Barto：Reinforcement Learning, The MIT Press, pp 135-150,1998

[2] C.H.J.C,Watkins：Q-Learning,Machine Learning 8, pp 279-292,1992

[3] 畝見　達雄：強化学習、 人工知能学会誌、 vol 9 no 6 pp 835-850

[4] 浅田　稔：強化学習の実ロボットへの応用とその課題、人工知能学会誌、 vol 12 no 9　pp 831-835

[5] 木村　元：部分観測マルコフ過程決定下での強化学習、人工知能学会誌、 vol 9 no 6　pp 822-829

[6] S.Russel and P.Norvig：エージェントアプローチ人工知能、共立出版、pp 601-628,1997