22E-01

実時間区間推定法によるQ学習の学習プランニング

岩田 一貴‡ 伊藤 暢浩‡‡ 山内 康一郎‡ 石井 直宏‡
名古屋工業大学 ‡知能情報システム学科 ‡‡電気情報工学科

1 はじめに

強化学習は、報酬を動機として試行錯誤を行ない適応現象を実現する学習手法である。教師あり学習手法とは違い、報酬という入力をもとに環境に対しての行動選択を行なう。よって常に正解を教えてくれるような教師を必要としない。また、経験を積み重ねながら行動に対する予測と結果の差を縮めてゆくので、時間をかけて変化する環境下でも学習を行なうことができる。このような学習手法は、例えばシステム障害時のような未知の環境下における行動選択に対して有効である。

強化学習の代表的な手法にQ学習 [1] がある。Q学習は実装が平易で遅延報酬に対して効果的なアルゴリズムだが、学習途中でのQ値の実用性、探査と知識利用のトレードオフ、探査の継続性などにおいての問題点がある。本稿では、Q学習とその問題点について述べ、その解決手法として実時間区間推定法を提案する。

2 Q学習

強化学習において学習者のとる行動は、即時に得られる報酬 (即時報酬) の決定だけでなく次状態も決定する。よって行動決定には、即時報酬と次状態以後から得られる報酬 (遅延報酬) が考慮されなければならない。そのため、学習者は即時報酬のみでなく遅延報酬からも学習できなければならない。

Q学習は、実装が平易で遅延報酬に対して効果的なアルゴリズムとして知られている。Q学習は、学習者が環境と相互作用を繰り返して、最適な行動-状態価値関数Qを見い出すことで、最適な行動選択戦略を求める。また、十分に試行を繰り返せば、Q値が既知のQ空間の最適値に収束することが証明されている [1]。しかし、Q学習には以下のような問題がある。

2.1 学習途中でのQ値の実用性

一般にQ値は漸近的に収束していく。学習の途中段階ではQ値の実用性は低い。Q値が有用になるには、収束を待たなければならない。そのために、Q値が最適に近い値になるまでの収束速度が問題となる。

2.2 探査と知識利用のトレードオフ

Q値が収束することとQ値が最適値に収束することとは無関係である。Q値は、既知のQ空間に真の最適値がなければ、局所的な最適値に収束してしまう。探査を積極的に行なえば、既知のQ空間を広げることができるが、多くの無駄な行動をとることになり、収束までに多くの試行回数と時間を必要としてしまう。つまり、既知のQ空間を広げること (探査) とQ値を収束させること (知識利用) はトレードオフの関係にある。Q学習では、このトレードオフをうまく調節しなければならない。

2.3 探査の継続性

学習が進行すると、多くのアルゴリズムでは、探査がほとんど行なわれないようになる。このために学習進行後の環境の変化に対応できない。動的な環境下では、収束後も継続して探査が続けられなければならない。

3 実時間区間推定法

ここでは、問題領域を離散マルコフ過程に限定する。離散マルコフ過程では、状態と行動は離散値として与えられる。ある環境下で各行動をとった場合の状態遷移は決定的であるとは限らないので、状態遷移は非決定的である。

最適な行動をとるためには、各行動をとった場合の遷移確率とその即時報酬値が明確になっていなければならない。そのために、実時間区間推定法では各行動に対する評価値と行動選択戦略を以下のように定義する。

3.1 行動価値評価

図 3.1.1 のようにある状態 S_0 からある行動 a によって互いに排反な状態 $(S_1, S_2, \cdots, S_k)$ のうちどれかに遷移し、その遷移確率を $(p_1, p_2, \cdots, p_k)$ とする離散マルコフ過程モデルを考える。

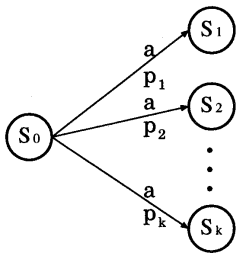

図 3.1.1 離散マルコフ過程モデル

Real Time Interval Estimation Method for Planning of Q-learning
Kazunori Iwata‡,Nobuhiro Ito‡‡,Kouichiro Yamauchi‡,Naohiro Ishii‡
Dept. of ‡Intelligence and Computer Science, ‡‡Electrical and Computer Engineering,Nagoya Institute of Technology
Gokiso-cho,Syouwa-ku,Nagoya 466-8555

この試行を N 回繰り返すとき、状態 $S_i(1 \leq i \leq k)$ へ遷移する回数を X_i とすれば、$(X_1, X_2, \cdots, X_k)$ の分布は多項分布 $M(N, p_1, p_2, \cdots, p_k)$ に従う。また、X_i の分布は二項分布 $Bin(N, p_i)$ に従う (証明略)。

行動 a を n 回試行した結果、$(S_1, S_2, \cdots, S_k)$ へ遷移した回数が $(x_1, x_2, \cdots, x_k)$ となった場合を考える。

実時間区間推定法では、行動 a についての遷移がどの程度明確になっているのかを示す尺度を行動発展レベルとして以下の表 3.1.1 のように定義する。

表 3.1.1 行動発展レベルと条件

レベル	条件
LEVEL1	n $\leq$ *enough trial*
LEVEL2	*enough trial* $<$ n

enough trial は定数で、$15k$ が目安である。

各行動発展レベルでの p_i の $100(1-\alpha)\%(\alpha =$ 有意水準) の信頼区間の上限値 P_i^U は F 分布・正規分布で近似して表 3.1.2 のように計算される [2]。

表 3.1.2 各レベルでの信頼上限値

レベル	信頼上限値 P_i^U
LEVEL1	$\frac{(x_i+1)F_{\frac{\alpha}{2}}(2(x_i+1),2(n-x_i))}{(n-x_i)+(x_i+1)F_{\frac{\alpha}{2}}(2(x_i+1),2(n-x_i))}$
LEVEL2	$\bar{p_i} + z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p_i}(1-\bar{p_i})}{n}} \quad (\bar{p_i} = \frac{x_i}{n})$

以上より、ある状態 S_0 における行動 a の行動選択評価値 V_a を以下のように定義する。

$$V_a = \sum_{i=1}^{k} P_i^U (r_i + \max_b Q(S_i, b))$$

r_i は、状態 S_i へ遷移した時に得られた即時報酬値の平均値とする。

3.2 行動選択戦略

アルゴリズムは以下のようになる。学習者は、現在の状態から以下の戦略で行動選択を行なう。

行動決定戦略

1. 各行動の行動発展レベルを計算する。
2. 最も低い行動発展レベルの行動を実行する。複数ある場合は、V 値が最大となる行動を選択する。

4 実験

4.1 問題設定

図 4.1.1 の離散マルコフ過程モデルで $\varepsilon-greedy$ による Q 学習と実時間区間推定法による Q 学習との比較実験を行なった。ただし、各パラメータは $\varepsilon = 0.2, \alpha = 0.05$ とした。

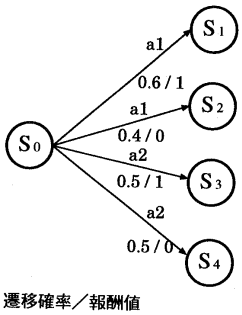

図 4.1.1 離散マルコフ過程モデル

4.2 実験結果

結果は、50 回の実験の平均値である。以下の表は行動 a1,a2 の選択回数と獲得報酬値を示す。

表 4.2.1 試行回数 50 回での結果

行動選択戦略	a1	a2	獲得報酬値
$\varepsilon-greedy$	30	20	16
実時間区間推定法	38	12	24

表 4.2.2 試行回数 1000 回での結果

行動選択戦略	a1	a2	獲得報酬値
$\varepsilon-greedy$	756	244	530
実時間区間推定法	959	41	592

4.3 考察

図 4.1.1 の離散マルコフ過程では、報酬が得られる確率が高い行動が a1 であることをより早く把握して a1 を多く選択するような行動選択戦略が好ましい。

表 4.2.1 および表 4.2.2 の結果から実時間区間推定法による Q 学習は、$\varepsilon-greedy$ による Q 学習と比較して、早い段階で環境を正確に把握しており、経験が比較的少ない段階においても適度に優れた行動選択を行なっている。また、$\varepsilon-greedy$ よりも無駄な探査が少なく、探査と知識利用のトレードオフをうまく調整していることがわかる。

5 まとめ

実時間区間推定法では、どの程度行動についての遷移が明確になっているのかを表すのに行動発展レベルを導入した。これにより、試行回数に応じた適切な評価式を用いることができ、さらにその時点での Q 値の実用性をはかることができる。また、行動発展レベルと行動評価値に基づいて決定される行動決定戦略は、探査と知識利用のトレードオフをうまく調節できることを示した。

参考文献

[1] Watkins, C.J.C.H., Dayan, P. Technical Note : Q-learning,Machine Learning 8, pp.55-68. (1992)

[2] 鈴木榮一・小林三郎・谷重雄 共著 統計学概論 地人書館