2K-04　POMDPs環境下での状態の価値を用いた学習法の提案*

後藤 亮　　松尾啓志†
名古屋工業大学電気情報工学科‡

1　はじめに

現在、機械制御を始めとする様々な分野において、その制御規則獲得のための学習法として強化学習が注目されている。強化学習とは教師なし学習の一種であり、エージェントは環境に対する知識を予め持たず、環境から与えられる報酬と観測という情報を手がかりに環境に適応する学習法である。そのためエージェントは未知の環境中を試行錯誤しながら行動することを通じて学習する。

近年の強化学習研究では、部分観測マルコフ決定過程 (POMDP) を対象としたものが多くなっている。POMDP は、不完全知覚状態、すなわち実際には環境の異なる状態がエージェントにとって同一のものとして知覚される状態を有しており、マルコフ決定過程 (MDP) を対象とした従来の学習法では、このような環境に適応することはできない。

POMDP を対象にした学習法の1つに確率的傾斜法 [木村 96] がある。確率的傾斜法とは平均報酬をもっとも大きくする方向へと政策を逐次的に改善する学習法である。確率的傾斜法の利点はメモリレスである点や計算コストのかかる処理が不要な点など多々ある。しかし、POMDP の基礎となる MDP の推定などは行わずに、報酬情報のみから学習するために、環境によっては最適解からはほど遠い解しか得られない場合がある。そこで、本研究では状態の価値を考慮する学習法を提案する。

2　環境

MDP の環境は最も基本的な環境で、状態遷移図で表すことができる。MDP は不完全知覚状態を持たないため、観測から状態を一意に定めることができる。一方、POMDP の環境は不完全知覚状態を有する。POMDP には図 1(a) のように、基礎となる MDP が存在しており、この基礎となる MDP 中の異なる複数の状態が同一の観測を持っている。図 1 では、異なる状態 S_1, S_2 が同一の観測 X_1 を持つため、エージェントには図 1(b) のような POMDP 環境として認識される。

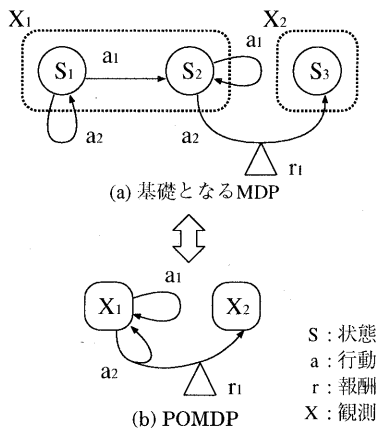


図 1: POMDP 環境

ここで、簡単な例として図 2(a) のような迷路の環境を考える。この迷路環境ではエージェントは始点 (S) から出発し、終点 (G) にたどり着くまで行動を繰り返し、終点にたどり着いたときに初めて報酬を得る。ここで、観測をエージェントの現在の位置のグローバルな座標とすれば迷路は MDP となるが、観測がエージェントの周囲 8 マスのパターンとなった場合、迷路は POMDP となる。図 2(a) の迷路では、終点にたどり着くための最適行動が、状態 A では右、状態 B では左と異なっているため、状態 A と B は異なる状態として認識されるべきである。しかし、図 2(b) からもわかるように、周囲 8 マスがまったく同じパターンであるため状態 A と B は同一視される。このため、観測からだけでは最適行動が定められず、従来手法をそのまま適用したのではエージェントの動作は安定しない。

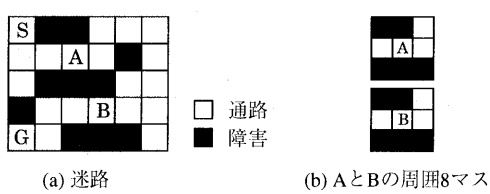


図 2: 不完全知覚が生じる迷路

*Learning Method with State Values in Partially Observable Markov Decision Processes
†Ryo Goto, Hiroshi Matsuo
‡Department of Electrical and Computer Engineering, Nagoya Institute of Technology

3 Q-learning

MDP を対象とした学習法に Q-learning [Watkins 92] がある。Q-learning とは、状態と行動の対に対して、その重みを Q 値として評価する学習法である。現在の状態が s_1 で行動 a_1 をとり s_2 に遷移し報酬 r を得た場合、Q 値は次式に従って更新される。

$$Q(s_1,a_1) \leftarrow (1-\alpha)Q(s_1,a_1) + \alpha(r + \gamma V(s_2)) \quad (1)$$

$$\left(\begin{array}{c} V(s) = \max_{a \in A} Q(s,a) \\ \alpha\text{: 学習率}, \gamma\text{: 割引率} \end{array} \right)$$

MDP では、最終的に Q 値が収束すると最大の Q 値を持つ行動の選択が最適政策となる。

4 提案手法

Q 値の更新式 (1) を解くことによって、収束後は、

$$Q(s_1,a_1) = r + \gamma V(s_2) \text{ 即ち } V(s_2) = \frac{Q(s_1,a_1) - r}{\gamma}$$

となることがわかる。Q 値がある程度理想的な値に収束したものと考えると、$Q(s_1,a_1)$ と r から状態 s_2 の価値 $V(s_2)$ が予測できる。このように、前の状態の情報から次の状態の価値の予測をすることによって、同一視された複数の状態を分離することを試みる。

まず、エージェント内部で 1 つの観測に対して複数の状態を予め用意しておく。これは、POMDP の基礎となる MDP 上での複数の状態が、1 つの観測の中に含まれている可能性があるためである。そして次の 1 ~3 を繰り返す。

1. 現在の状態の Q 値から次状態の価値を予測する
2. 環境から与えられた観測の中から予測値に近い価値をもつ状態を選び出す
3. 選出した状態を遷移先状態として、通常の Q-learning と同様に Q 値の更新を行う

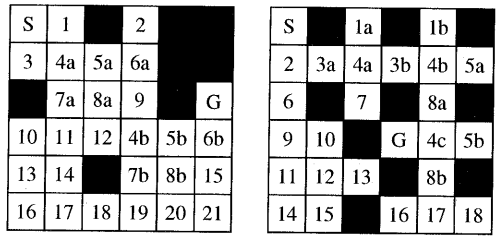
このステップを繰り返すことで、同一視された状態の分離と学習を試みている。

5 実験

提案手法を評価するために、Q-learning と確率的傾斜法と提案手法の比較実験を行った。

5.1 内容

図 3 のような 2 種類の迷路を用い、観測をエージェントの周囲 8 マスとして 100 回の実験を行い、終点到達までにかかったステップ数の平均値と最小値を 3 つの手法で比較した。図 3 の迷路は (a)(b) ともに POMDP 環境となり、同一視される状態を同じ数字で示した (8a と 8b など)。同一の観測となる状態の最大個数は、迷路 (a) では 2 つ (4a と 4b、5a と 5b など)、迷路 (b) では 3 つ (4a,4b,4c) となっている。提案手法では予め各観測内に 3 つの状態を作成した。パラメータや学習回数は結果が最良になるよう調整した。



迷路 (a)
最適ステップ数 : 9

迷路 (b)
最適ステップ数 : 8

図 3: 実験に用いた迷路

5.2 結果と考察

表 1: 実験結果

	迷路 (a)		迷路 (b)	
	平均	最小	平均	最小
Q-learning	33.74	11	195.58	22
確率的傾斜法	13.36	9	30.96	8
提案手法	12.85	9	11.08	8

Q-learning では迷路 (a)(b) ともに最適ステップ数は一度も得られず、平均的にも最適解からは遠い結果しか得られなかった。確率的傾斜法と提案手法を比較すると、両手法とも最小値には最適ステップ数が得られたが、迷路 (b) のステップ数の平均値に着目すると、提案手法は確率的傾斜法に比べても良好な結果が得られたことがわかる。これは、提案手法では POMDP の基礎となる MDP の構造を考慮しているためであると考えられる。

6 まとめ

部分観測マルコフ決定過程環境下において状態の価値を考慮する学習法を提案し、実験によりその有効性を確認した。今後の課題としては、現在各観測内に静的に作成している状態を環境に適応して動的に生成するように改善することや、挙動を詳しく解析することなどが挙げられる。

参考文献

[1] 木村元, 山村雅幸, 小林重信 : 部分観測マルコフ決定過程下での強化学習 : 確率的傾斜法による接近, 人工知能学会誌, Vol.11, No.5, pp.761-768(1996).

[2] Watkins, C.J.C.H. and Dayan, P.: Technical Note: Q-Learning, *Machine Learning*, Vol.8, No.3, pp.279-292(1992).