2004 年 12 月 8 日

知能型システム論: 強化学習とマルコフ決定問題

喜多 一

1 強化学習

環境の中で動作するシステムがスカラーの報酬あるいは罰として，その行動の良否を評価を得るものとする．このとき，この評価を最大化するように行動の獲得を進める機械学習を強化学習 (reinforcement learning) と呼ぶ．また，与えられる評価を強化信号 (reinforcement signal) と呼ぶ．

このような学習方式を考えるに際しては，以下のような幾つかの問題がある：

1. 行動に対する評価には時間的遅れがある．これへの対策.
2. 最適な行動を模索するための効果的な試行錯誤の実施.
3. 行動の決定に際しての環境の知覚の不完全性への対応.
4. シミュレーションと実行動の差異.

以下では上記の 1, 2 に焦点を当てつつ，強化学習の代表的手法である Q-学習を紹介する準備として，マルコフ決定問題とその解法である Value Iteration 法について解説する.

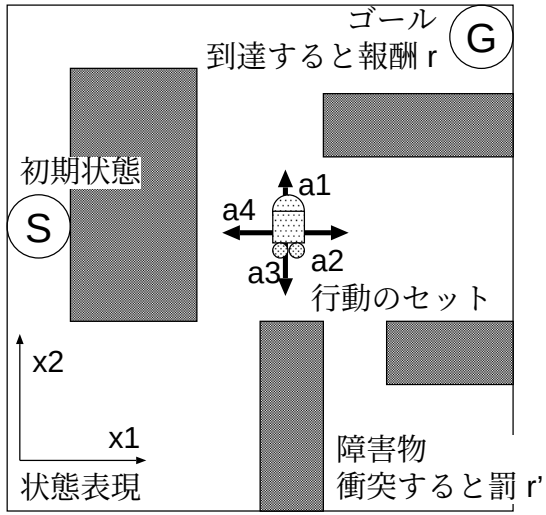

図 1 強化学習

2　マルコフ決定問題

強化学習における最大の問題は行動に対する評価の時間的遅れへの対策である．すなわち，一連の系列として取られた行動の結果として得られた評価値を時間を遡って個々の行動の決定に関連付けなければならない．そのための枠組を与えるモデルの一つがマルコフ決定問題 (Markov decision problem, MDP) である．

いま，環境の中で

- 離散時間 $t = 1, 2, 3 \cdots$ で動作するシステムを考える．
- システムは有限の「状態」 $x_i \in X$, ($|X|$ は有限) を取るものとし，
- 各状態で有限の選択肢から「行動」 $a_j \in A$, ($|A|$ は有限) をその状態に応じて取ることができるものとする．
- システムは現在の状態 x_i と取った行動 a_j に応じて確率 p_{ijk} で次の状態 $x_k \in X$ に遷移するとともに，
- 強化信号 $r_{ijk} \in R$ を得る．

システムの評価は強化信号の「割引かれた累積値の期待値」

$$V = \mathrm{E}\left(r(1) + \gamma^1 r(2) + \gamma^2 r(3) + \cdots\right) = \mathrm{E}\left(\sum_{t=1}^{\infty} \gamma^{(t-1)} r(t)\right) \tag{1}$$

として与える．ここで $0 < \gamma < 1$ は割引き率と呼ばれる定数で，1 ステップ未来の強

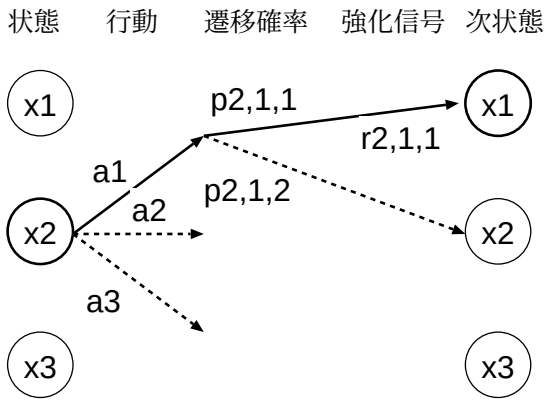

図 2　マルコフ決定問題

化信号を γ だけ割引いて現在の評価とするためのものである．また $r(t) \in R$ はシステムが得る強化信号 r_{ijk} を時間系列として表したものである．なお，記法 E() は状態遷移確率に関する期待値を表す．

システムが状態 x_i に応じて行動 a_j を決定する関数 $\mu : X \to A$ を「政策」と呼ぶ．マルコフ決定問題 (Markov decision problem, MDP) とは

> MDP: 与えられた初期状態 $x(0) \in X$ に対して，評価値 V を最大にする政策 μ^* を求めることである．

すなわち

$$\text{find} \quad \mu^* \quad \text{such that} \quad \mu^* = \arg\max_{\mu} V(x(0, \mu) \tag{2}$$

として表される．

MDP の定式化を用いることにより行動の系列と強化信号 (の系列) が明確に関係づけられ，強化学習の問題が MDP の解法として位置付けられる．ただし，MDP では環境の知覚については状態 x_i が曖昧さなく得られるという意味での理想化が行われている．

割引き率の導入により強化信号 r が有界ならば，時間ステップが無限になっても評価値 V が有界となる．

3　動的計画法

MDP の解法の基礎となるものが動的計画法 (dynamic programming) である．動的計画法の考え方の基本は

> **最適性の原理**：未来の行動が最適になされるものと仮定し，その評価を利用して現在の行動を最適に決定すればその行動は最適行動にほかならない．

というものである．

この原理は最適制御理論や有効グラフの最短経路を求めるアルゴリズムなどにも利用されており，多段階の最適決定問題の基礎的な考え方である．時間ステップが有限ならば動的計画法は最終段階から逐次時間を遡ることで適用できる．しかし，先に定式化したマルコフ決定問題は時間を無限大まで考えるため，動的計画法を用いて解を得るには工夫が必要である．

動的計画法の考え方に従えば MDP については，

- いま状態 x_i にいて，行動 a_j を決定しなければならないとしよう．
- これにより状態は確率的に x_k に遷移する訳であるが，
- x_k への遷移後は最適な政策 μ^* のもとで未来の決定がなされると想定しよう [*]．
- そして，その場合の状態 x_k の評価値を仮に $V^*(x_k)$ とする．

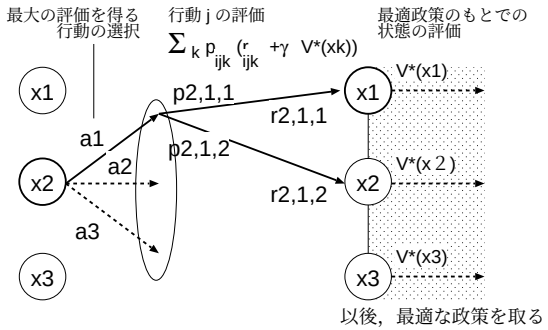

図 3　動的計画法

[*] それは分からないではないか，と怒ってはいけない．作業仮説であり，とりあえず未来を楽観しして，現在を考えるという方針である．

この想定のもとでは，状態 x_i における最適な決定は，行動 a_j を取ったときの評価値の期待値

$$\sum_k p_{ijk}(r_{ijk} + \gamma V^*(x_k)) \tag{3}$$

を最大にする行動 $a_j \in A$ を選ぶことである．したがって，最適な評価値 V^* については

$$V^*(x_i) = \max_{a_j} \sum_k p_{ijk}(r_{ijk} + \gamma V^*(x_k)) \tag{4}$$

が成り立つ．

残念ながら (4) 式は両辺に V^* が現れ，かつ右辺には最大値を求めるという演算が含まれるため，そのまま解くことはできない．

これを求める具体的なアルゴリズムは幾つか知られているが，以下では強化学習のアルゴリズムの代表例である Q-学習法と関連の深い Value Iteration 法を紹介する．

4 Value Iteration 法

Value Iteration 法は (4) 式を V^* に関する漸化式として考え，繰り返し計算により，その収束値として解 V^* を求めるものである．すなわち

適当な仮定のもとでの収束性は保障されている．

1. 初期の評価値 $V^0(x_i), x_i \in X$ を適当に設定する．
2. 繰り返しのカウンタ $l = 0$ とする．
3. 以下の漸化式で $V^l(x_i), x_i \in X$ を求める．

$$V^{l+1}(x_i) = \max_{a_j} \sum_k p_{ijk}(r_{ijk} + \gamma V^l(x_k)), \quad \text{for} \quad x_i \in X \tag{5}$$

4. 評価値 V^l が十分収束すれば終了．そうでなければ $l = l + 1$ として 3. にもどる．

参考文献

[1] D. Bertsekas: Dynamic Programming, Prentice-Hall (1987)

[2] D. Bertsekas and J. Tsitsiklis: Neuro-Dynamic Programming, Athena Scientific (1996)

[3] R. Sutton and A. Barto (三上，皆川訳)：強化学習, 森北出版 (2000)

演習

図 4 のような 6 状態のマルコフ決定問題を考える.

- 各状態で取り得る行動は右に遷移する (R) か，左に遷移する (L) かの 2 つであり,
- 状態遷移は確定的であるものとする.
- ただし，両端の状態 (x_1, x_6) では，行動によらずその状態に留まるものとする.
- 強化信号は状態遷移 $x_2 \to x_1$ および $x_5 \to x_6$ に際して，それぞれ報酬 1 および 2 が得られるものとする．それ以外では強化信号は入らない.
- 割引き率は $\gamma = 0.9$ とする.

この問題に対して状態の評価値 V の初期値を 0 として，Value Iteration アルゴリズムを適用し，各状態において最適な行動を求めよ.

ヒント　確定的な状態遷移を考えており，行動は R と L の 2 種類であるから (5) 式は

$$V^{l+1}(x_i) = \max\left(r_{i\mathrm{R}k} + \gamma V^l(x_k), r_{i\mathrm{L}k'} + \gamma V^l(x'_k)\right)$$

である．ここで k は各状態での右遷移の行き先， k' は左遷移の行き先である.

また強化信号 $r_{i\mathrm{R}k}, r_{i\mathrm{L}k'}$ は状態遷移が $x_2 \to x_1$ および $x_5 \to x_6$ の場合にそれぞれ 1 および 2 であり，それ以外は 0 である.

Step		$V(x_1)$	$V(x_2)$	$V(x_3)$	$V(x_4)$	$V(x_5)$	$V(x_6)$
0	初期評価値	0	0	0	0	0	0
1	行動 L の評価値	0					0
	行動 R の評価値	0					0
	評価値 V	0					0
2	行動 L の評価値	0					0
	行動 R の評価値	0					0
	評価値 V	0					0
3	行動 L の評価値	0					0
	行動 R の評価値	0					0
	評価値 V	0					0
4	行動 L の評価値	0					0
	行動 R の評価値	0					0
	評価値 V	0					0
5	行動 L の評価値	0					0
	行動 R の評価値	0					0
	評価値 V	0					0

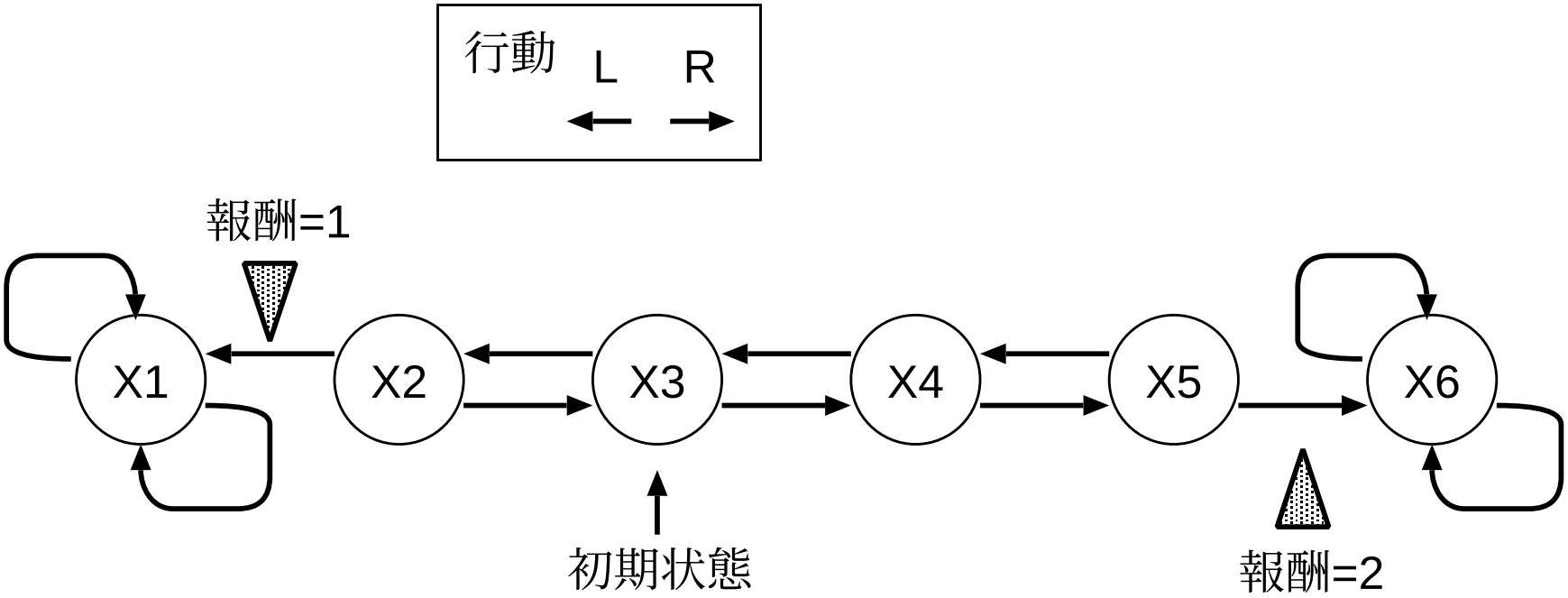

図 4　6-状態マルコフ決定問題