2004 年 12 月 15 日

知能型システム論: Q-学習法

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.

以下では強化学習の代表的手法である Q-学習を紹介する.

2 マルコフ決定問題 (復習)

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} を時系列として表したものである．なお，記法 $\mathrm{E}()$ 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.

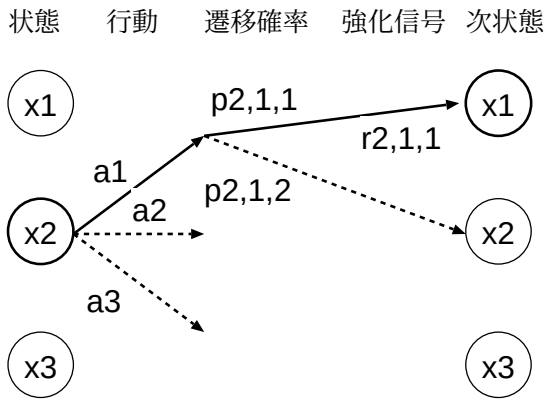

図 1　マルコフ決定問題

3　Value Iteration 法 (復習)

MDP の解法として動的計画法の考え方に基づき，最適政策の評価値 V^* を繰り返し計算により求める Value Iteration 法がある．

この方法は以下の繰り返し計算により，その収束値として解 V^* を求める：

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．

$$V^{l+1}(x_i) = \max_{a_j} \sum_k p_{ijk}(r_{ijk} + \gamma V^l(x_k)), \quad \text{for} \quad x_i \in X \tag{3}$$

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．

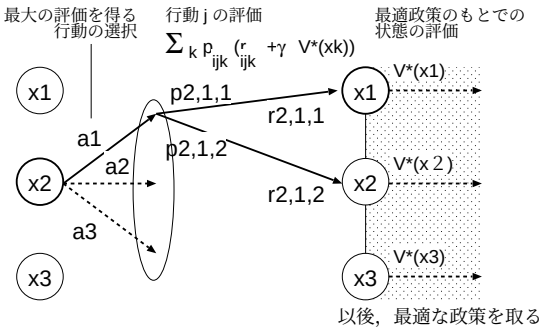

図 2　動的計画法

4 Q-学習法

4.1 学習法としての VI 法の問題

VI 法は MDP のすべてのパラメータが既知であるとして解を求めている．しかしながら「強化学習」として問題を扱う場合にはシステムのパラメータには未知のもの (例えば状態遷移確率 p_{ijk}) が含まれることが前提となる[*1]．Q-学習法 (Q-learning) は状態遷移確率 p_{ijk} や強化信号 r_{ijk} が未知の場合に，システムが試行錯誤を自ら繰り返すことにより評価値を最適化する政策 μ^* を獲得する方法である．

4.2 Q 値

Q-学習法では状態 x_i, 行動 a_j の関数 (テーブル) として $Q(x_i, a_j)$ という値 (Q 値) を考える．$Q(x_i, a_j)$ は

> 「現在，状態 x_i にあり，そのとき (任意の) 行動 a_j を取るものとし，それ以降は最適政策をとると想定した場合の評価値の推定値」

[*1] そうでなければ「学習」を考える意味がない

である．学習が収束した段階では最適な評価値 V^* と獲得した Q 値 (Q^* とする) の間には

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

$$= \sum_k p_{ijk} \left(r_{ijk} + \gamma \max_{j'} Q^*(x_k, a'_j) \right) \tag{5}$$

$$V^*(x_i) = \max_{a_j} Q^*(x_i, a_j) \tag{6}$$

という関係が成立する．

4.3 行動決定ルール

状態遷移確率 p_{ijk} や強化信号 r_{ijk} が未知であれば，各状態で可能なすべての行動を (十分に多くの回数) 試みなければ最適政策は獲得できない．

一方，限られた学習の機会を有効に利用するためには，より高い評価の得られる可能性のある行動をより重点的に試みることが考えられる．

これらを具体化した行動決定ルールとして

ϵ-Greedy 法 ：$0 < \epsilon < 1$ をパラメータとし，確率 $(1 - \epsilon)$ で最大の Q 値を持つ行動を，確率 ϵ でその他の行動を (等確率で) ランダムに選択する手法である．

ソフトマックス法 ：$0 < \tau$ をパラメータとし，指数関数 $\exp(Q(x_i, a_j)/\tau)$ に比例し

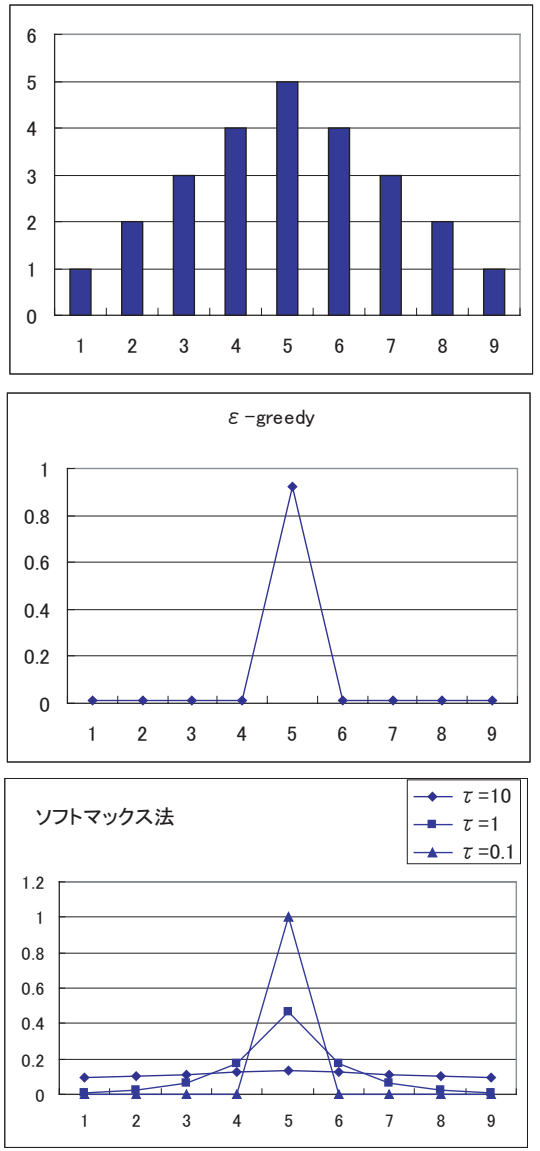

図 3　行動決定ルール

た確率でランダムに行動を選択する手法である[*2]．パラメータ τ が小さければ Q 値の大きい行動を重点化して選択し，τ が大きければすべて行動を似た確率で選択する．

なお，以下に述べる Q-学習法そのものの行動選択ルールに対する要請は緩やかであり，すべての状態においてすべての行動が十分に多くの回数選ばれるものであればよい．

4.4 平均値の逐次計算

VI 法を学習型にするために，まず準備として n 個のデータの平均値を求める計算法について考える．データ $x_1, x_2, \cdots x_n$ の平均は

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

で与えられるが，これは

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n-1} x_i + x_n\right) \tag{8}$$

$$= \frac{n-1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n-1} x_i\right) + \frac{1}{n}x_n \tag{9}$$

[*2] パラメータ τ は統計力学におけるボルツマン分布との対比で「温度」と呼ばれる．

と書ける．t 個までの平均値を $\bar{x}(t)$ と書くことにすれば逐次的に平均値を求める漸化式

$$\bar{x}(t+1) = \left(\frac{t}{t+1}\right)\bar{x}(t) + \left(\frac{1}{t+1}\right)x(t+1) \tag{10}$$

$$= (1-\alpha(t+1))\bar{x}(t) + \alpha(t+1)x(t+1) \tag{11}$$

が得られる．ここで $\alpha(t+1) = 1/(t+1)$ である．

4.5 Q 値の学習

時刻 t において状態 $x(t) = x_i$ で ε-greedy 法など何らかの方法で行動 $a(t) = a_j$ を取り，(確率的に) 状態が $x(t+1) = x_k$ に遷移し，その際，強化信号 $r(t)$ が得られたものとする．このとき VI 法のアルゴリズムから，これを p_{ijk}, r_{ijk} が未知の場合に学習で実行する方法を考える．

まず (3) 式の VI 法の更新式は以下のように書くことができる：

$$Q^{t+1}(x_i, a_j) = \sum_k p_{ijk}(r_{ijk} + \gamma V^t(x_k)) \tag{12}$$

$$V^{t+1}(x_i) = \max_{a_j} Q^{t+1}(x_i, a_j) \tag{13}$$

さらに，(12) 式に (1 ステップ前の)(13) 式を代入すると

$$Q^{t+1}(x_i, a_j) = \sum_k p_{ijk}(r_{ijk} + \gamma \max_{a_l} Q^t(x_k, a_l)) \tag{14}$$

となる．この式で期待値を確率 p_{ijk} を既知として求める変わりに実際に生じた状態遷移 $x(t) \to x(t+1)$ を用いて逐次的に平均を求める方法を用いて，Q 値を以下のように更新する

$$Q^{t+1}(x(t), a(t)) = (1-\alpha)Q^t(x(t), a(t)) + \alpha(r(t) + \gamma V(x(t+1))) \quad (15)$$

$$V^t(x(t+1)) = \max_{a_j} Q^t(x(t+1), a_j) \quad (16)$$

あるいは，これらを一括して

$$Q^{t+1}(x(t), a(t)) = (1-\alpha)Q^t(x(t), a(t)) + \alpha(r(t) + \gamma \max_{a_l} Q^t(x(t+1), a_l)) \quad (17)$$

を得る．ここで $0 < \alpha < 1$ は学習係数と呼ばれるパラメータであり，状態遷移や強化信号の不確実性に対して，その平均を学習により求めるために必要とするものである．先の平均値を求める議論からも推測されるように Q 値の収束を保証するためには α を徐々に小さくしてゆくことが要請される．

5 演習課題

5.1 問題

図 4 に示すような 8×8 マスの格子上の環境で学習エージェント (A) にスタート (S) からゴール (G) への経路を Q 学習法により獲得させる迷路学習課題の計算機実験を試みよ.

問題設定の詳細は以下の通りである：

状態 ：状態としてエージェントは現在の位置 (x, y) 座標を用いる.

初期状態 ：初期状態はスタート地点 $S = (s_x, s_y) = (0, 4)$ である.

行動 ：エージェントの取り得る行動は東西南北 4 方向への 1 マスの移動 $A = \{\mathrm{E}, \mathrm{W}, \mathrm{S}, \mathrm{N}\}$ である.

状態遷移 ：移動しようとする先によって以下のように状態遷移する:

- 現在位置がゴール G であれば，取った行動によらずスタート S に遷移する.
- もし，移動先が壁 (灰色の部分，8×8 の格子の外側への移動を含む) であれば，移動はできず，エージェントは現在の場所に留まる.
- それ以外であれば，移動しようとする先に遷移する.

強化信号 ：強化信号として以下の報酬，罰が与えられる：

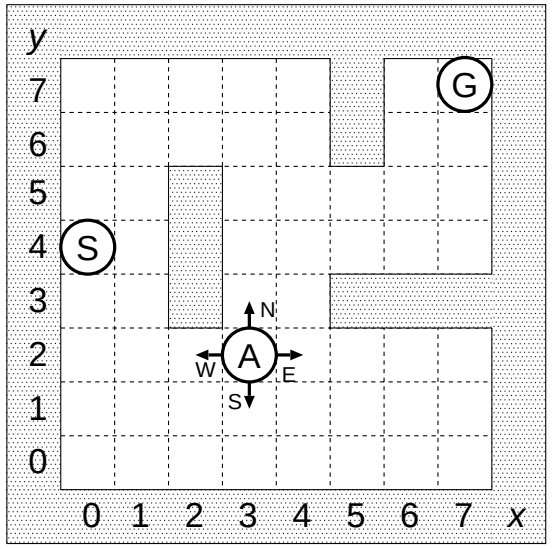

図 4　迷路学習課題

- ゴールにたどりついた場合に報酬として 1 が与えられる.
- 壁や外側への移動を試みた場合には罰として -0.1 が与えられる.
- それ以外の場合には強化信号は与えられない.

割引率 : $\gamma = 0.9$ とする.

なお, 実験に当たって必要に応じて設定を適宜調整してよい.

5.2 Q-学習法

簡単に Q-学習法のアルゴリズムを示しておく.

1. Q 値 (この例題では $Q(x, y, a)$, $x \in X = 0, \cdots, 7$, $y \in Y = 0, \cdots, 7$, $a \in A = \{\mathrm{E}, \mathrm{W}, \mathrm{S}, \mathrm{N}\}$ となる) を初期化 (例えば 0) にする.
2. 時刻 $t = 0$ とする.
3. エージェントを初期位置 S に置く. $x = s_x, y = s_y$
4. 行動決定ルール (例えば ϵ-Greedy 法) により行動 a を決定する.
5. 状態遷移ルールに従い, 次状態 x', y' を得る.
6. 強化信号 r を得る.
7. Q 値を学習する.

$$Q(x, y, a) = (1 - \alpha)Q(x, y, a) + \alpha(r + \gamma \max_{a_l} Q(x', y', a_l))$$

8. 状態を更新する. $x = x', y = y'$

9. 時刻を 1 つ増やす．$t = t + 1$
10. 適当な条件で終了を判定する．終了でなければステップ 4. にもどる．

なお，パラメータ ϵ, α などの値や終了判定条件は適宜検討すること．

学習の進捗状況は各マス目において Q 値を図 2 のようなレーダーチャートで描くと分かりやすい．

5.3 拡張

余力があれば課題をさまざまに調整して学習を試みよ．例えば以下のようなことを検討せよ．

1. ゴールでの Q 値は学習しないようにするとどうなるか．
2. 壁に当たったときに罰を与えないようにするとどうなるか．
3. ゴールにたどり着いたあとに状態がスタートにもどらずに，ランダムな位置に遷移するようにするとどうなるか．
4. 状態表現として (x, y) の変わりに隣接したマスが壁やゴールであるかどうかという表現を用いるとどうなるか．このような表現はマルコフ決定問題が想定している「状態」と言えるか．

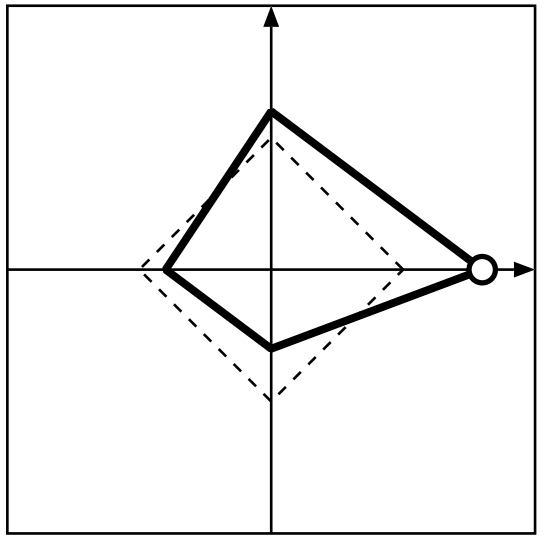

図 5　Q 値のレーダーチャートによる表現．実線は各行動 (移動方向) に対する Q 値を，点線は Q 値が 0 であることを表す．白丸は最大の Q 値を持つ行動．

5.4 提出方法

提出日 1 月 12 日（水）の授業開始時

書式 A4 判用紙を用いること

提出内容 以下の内容を含むこと

- 提出者の所属，氏名，学籍番号
- プログラムのソースコード（GUI などを含んで長い場合は学習アルゴリズムの部分のみでよい) とその説明.
- 実験の設定（パラメータの値など）と実験結果
- 考察

注意 レポート作成に関して以下の点に注意すること.

- 著しく類似したソースコードのレポートが見つかった場合は採点しない.
- 引用した文献や Web サイト（この授業の資料を除く）は参考文献として明記すること．また，レポート作成に際して，他の人の協力，助言を得た場合は謝辞として明記すること.