強化学習アルゴリズムとは
～試行錯誤的なアプローチ～

星野 孝総（高知工科大学）

平成 27 年 10 月 27 日

1　はじめに

学習アルゴリズムの研究は，Wolf と Hoff のデルタ則に端を発し，さまざまな学習則が提案され，多くの成果を上げてきた．また，試行錯誤によって学習を進める教師無し学習の代表的な手法として強化学習が提案され，自立エージェントの学習アルゴリズムとして研究が進んでいる．本稿では，強化学習の基本理論を解説する．これらの解説が強化学習研究の参考になることを期待する．

2　強化学習

強化学習は，報酬・罰を手がかりとして環境に適した行動を強化する学習法である．その由来は心理学の「パブロフの犬」を基本としている．また人工知能の早期において，強化学習は機械学習の一種であった．現在では，自立的に学習するエージェントの学習手法として研究されている．しかし，パラメータに敏感であることや，学習に時間が掛かるなどの問題点がある．強化学習の特徴は不確実性や報酬・罰遅れを伴った情報でも学習できることである．強化学習は図 1 に示すように，状態認識器，行動選択器，学習器の三つのユニットから構成されている．状態認識器は状態を認識して，政策候補の集合を生成し，行動選択器に送る．行動選択器は，状態認識器から送られた政策候補の集合から評価値の大きい行動を選択して環境に出力する．この政策により状態が遷移し，遷移先状態が報酬・罰の条件を満たしているとき環境は報酬・罰を学習器に与える．学習器は，報酬・罰に従って政策に関する評価値を変更する．

強化学習での，報酬 (reward)，罰 (penalty) は政策に対して遅れがあり，得られる条件は遷移先状態によって決定される．したがって，学習の目的はその報酬を多く得ることであり，言いかえれば，時間軸上の未来に対する報酬の総和を最大にすることになる．図 2 の場合，破線部分の総和は+1 となる．この総和を遷移先状態の評価値とし，(1) 式で与える．ここで，V は状態の評価値である．γ は割引率，r_i は時刻 t で得られる報酬である．r_i が正の値を取る時に報酬となり，負の値を取る時に罰となる．政策決定では，報酬が効率良く得るために，評価値 V が大きい状態に遷移する政策を選択する．

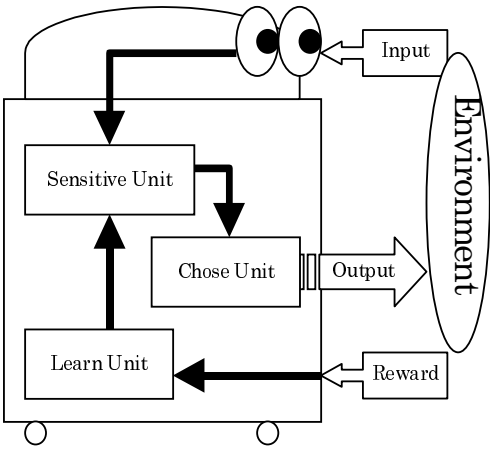

図 1: 強化学習

$$V = \sum_{i=t}^{\infty} \gamma^{i-t} r_i \tag{1}$$

学習過程では，(1) 式を実際に計算することはできない．そこで，学習器では，(2) 式に示すように，離散時間における評価値を更新する．ここで V_t は，時間 t における評価値である．f は強化関数と呼ばれ，時刻 t における報酬 r_t に関する関数である．報酬 r_t は報酬を受けた時は正の値をとり，罰を受けた時は負の値をとる．したがって，評価値は報酬を得られた時に正の方向に更新され，罰を得た時に負の方向に更新される．また，強化学習では図 3 に示すように新たな政策を開始してから報酬を受けるまでをエピソードと言い，離散時間のことをステップと言う．

$$V_t \longleftarrow V_t + f(r_t) \tag{2}$$

教師付き学習の場合は，強化関数に相当する関数に教師データと出力との差を使用する．具体的には，教師データと出力の差を縮めるように評価値を更新する．つまり教師データを目的値とするデルタ則になる．強

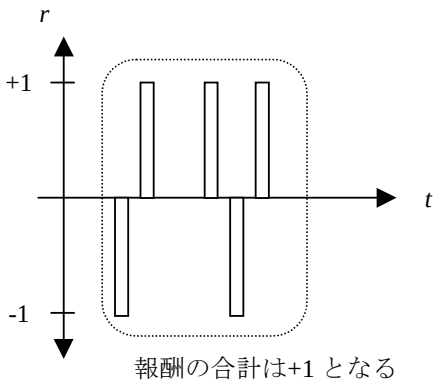

図 2: 報酬の総和

化学習では教師データが無く，報酬の総和を最大にする事を目的としているため，強化関数を報酬 r_t の関数になる．

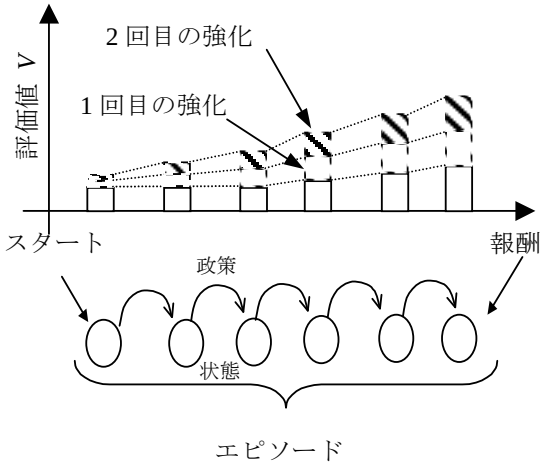

図 3: エピソードと強化

3　経験強化型強化学習

報酬が得られた経験に対してのみ強化する方法を経験強化型強化学習という．経験強化型の代表的な手法である ProfitSharing 法 (報酬割り当て法) の強化関数を (3) 式に示す．r は報酬であり，γ は割引率である．T は報酬 r を得た時刻，t は過去の時間である．これは，過去に経験した評価値を全て強化するため報酬獲得を重視した手法である．

$$f(r) = r\gamma^{T-t} \text{ ただし，} 0 < \gamma < 1 \qquad (3)$$

γ が大きい場合，過去に報酬獲得した全ての評価値を強化するため，図 4 に示すように無駄な経路の評価値まで強化するため，最適政策を得ることができない．反対に小さい場合はエピソード初期の評価値の強化が小さいためランダム探索が終了しない．特にエピソード長が変化する場合，エピソード初期の評価値の強化が安定せず学習が進まない．

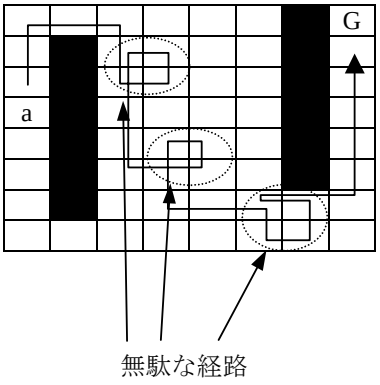

図 4: 報酬獲得

4　環境同定型強化学習

環境同定型は，現在と過去の評価値の差を強化関数に用いる手法である．つまり，図 5 に示すように過去から未来にたいする報酬の見積もり値を算出することになる．この評価値の差を TD-error と呼び (4) 式で与える．

$$\text{TD-error} = \gamma V_{t+1} - V_t \qquad (4)$$

この TD-error を用いた強化学習法を TD 法 (Tempo-

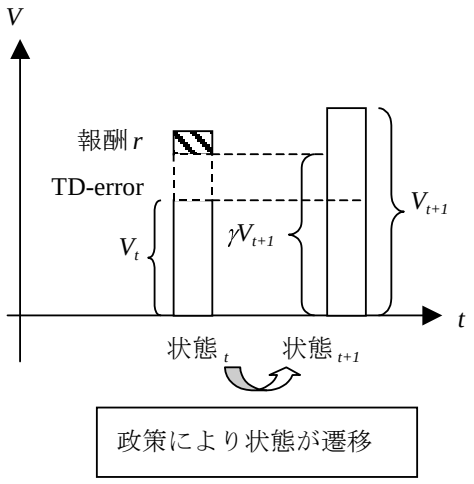

図 5: 報酬見積もり

ral Differential Method) と言い，(5) 式のような更新式を用いて学習を用いる．α は学習率であり，学習の速度を決定するパラメータである．また，γ は割引率であり，行動連鎖のつながりを示している．α と γ は，それぞれ [0,1] の値をとる．

$$V_t \longleftarrow V_t + \alpha(r + \gamma V_{t+1} - V_t)) \tag{5}$$

5　決定的政策決定と確率的政策決定

環境のモデルがある場合の遷移先状態は，決定的に決定でき，遷移先状態に対する評価値を政策に反映させれば良い．この政策決定を決定的政策決定という．(6) 式に決定的政策決定に用いるルールを示す．このルールでは，状態が a の時に評価値が V になることを示している．

$$\text{if 状態 is } a \text{ then 評価値 is } V \tag{6}$$

反対に，モデルが無い場合の遷移先状態は，決定的に決定できない，そこで状態と政策を一組として評価し，その政策を政策決定に反映させる．モデルが無いため，政策候補の適用が常に可能であるとは限らず，確率的である．そのためこの政策決定を確率的政策決定と言い，政策決定は確率的に行う必要がある．確率的政策決定で用いるルールは (7) 式のようになる．このルールでは，状態が a で政策 b を行った時に評価値が V^* になることを示している．確率的政策決定では，先にも述べたように遷移先状態を算出する必要がなく，モデルを必要としない．このような強化学習法はモデルレス型と呼ばれている．

$$\text{if 状態 is } a \text{ and 政策 is } b \textit{ then } \text{評価値 is } V^* \tag{7}$$

6　ルールテーブル型とルール追加型

ルールテーブルとはルール数を固定して離散分割して用意しておく方法を言う．それに対し，ルール追加型は，過去に適用未経験のルールをルールベースに追加していく方法である．ルールテーブルの場合は，離散分割しているため学習が早く進む．しかし，分割個数が政策の信頼性に大きな影響を与える．また，ルール追加型は必要なルールを効率良く収集できる．しかし，ルール数が膨大になる可能性があり，学習が遅くなると言われている．

7　Q-learning

環境同定型強化学習の代表的手法である Q-learning は，TD 法を確率的政策決定に発展させた手法である．Q-learning では，モデルが無いため遷移先状態を計算できない．したがって，政策を決定できない場合であっても，存在する状態を Q 値から同定しなければならない．Q-learning は，ある状態 a における政策 b に対して報酬の見積もり値 $Q(a,b)$ を算出し，その報酬の見積もりの大きい政策を選択する手法である．また，Q 値の更新は，(8) 式に示すように現在の状態に対する報酬に基づいて更新項を計算する．

$$Q(a,b) = (1-\alpha)Q(a,b) + \alpha(r + \gamma Q^{max}) \\ (Q^{max} : \text{次の政策の最大 } Q \text{ 値}) \tag{8}$$

ここで，α は学習率である．環境同定型強化学習は報酬の与え方や，Q 値の環境同定の効率により学習特性が大きく変化する．Q 値は対象とする状態の次状態の最大 Q 値 (Q^{max}) によって更新される．エージェントは試行錯誤を繰り返し，Q 値を更新して学習をすすめる．Q 値は状態と政策を対に持つので，学習初期では未知状態に進入し，Q 値を構成する．

8　おわりに

本論文では，強化学習の研究で用いられている手法を解説した．実際に熟練者の技術を試行錯誤的に習得する事を考える時，動的環境に対する環境同定は，強化学習にとって重要な課題である．そこで，本論文が実用的な強化学習を研究するための参考になることを期待する．

参考文献

[1] 畝見達男:強化学習:人工知能学会誌，Vol.9，No.6，pp.40-46 (1994)

[2] 宮崎和光，山村雅幸，小林重信:エージェントの学習:人工知能学会誌,Vol.10,No.5,pp.682-689 (1995)

[3] 宮崎和光，山村雅幸，小林重信：強化学習における報酬割り当ての理論的考察：人工知能学会誌 Vol.9,No.4,pp.580-587 (1994)

[4] 宮崎和光，山村雅幸，小林重信：強化学習の特徴と発展の方向：システム/制御/情報,Vol.39,No.4,pp.191-195 (1995)