2A1-2

認知モデルとしての強化学習

Reinforcement Learnig for a Cognitive Model

岡田　浩之
OKADA, Hiroyuki

東海大学理学部情報数理学科
Department of Mathematical Sciences, School of Science, Tokai University

Reinforcement learning method is now being actively studied as a framework for autonomous learning because actions can be learned using only scalar evaluation values and without explicit training signal. To solve the problem of tradeoff between exploration and exploitation actions in reinforcement learning, the authors have proposed two-dimensional evaluation reinforcement learning, which distinguishes between reward and punishment evaluation forecasts. In the proposed method of reinforcement learning using the two dimensions of reward and punishment, a reinforcement signal dependent on the environment is distinguished into reward evaluation after successful action and punishment evaluation after an unsuccessful action.

1. はじめに

強化学習は対象となるシステムの明示的なモデルを持たず、自己の行動の結果に対して環境から得られた報酬だけを手がかりに、適切な行動戦略を試行錯誤的に獲得する学習のクラス一般を指す用語である。明示的に教示すること無しに、報酬というスカラーの評価値だけを利用しての学習が可能なため、自律的な学習の枠組みとして研究がさかんになっている。

本稿では、初めに強化学習の代表的な枠組みである Actor-Critc 手法に従い、アルゴリズムの定式化を行う。次いで、報酬を基にした学習の枠組みであるにも関わらず，従来の強化学習研究では環境から与えられる報酬の取扱いが曖昧であったことを指摘し強化学習において報酬を多元的に扱うことの意義を論じる。特に、動物心理学との対比で負の報酬が行動に与える影響の重要性を指摘し、報酬性評価と嫌悪性評価を区別して扱う二次元評価強化学習において、多次元評価の行動次元への反映原理を議論する。

2. 強化学習の定式化

2.1 Actor-Critic 手法

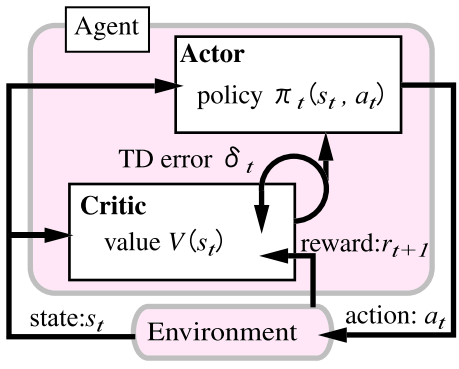

図 1: Actor-Critc 手法のアーキテクチャ

図 1 に Actor-Critic 手法の概要を示す。Critic は現在の環境の状態 (s_t) と環境から与えられる報酬 (r_{t+1}) から、状態に対する価値 ($V(s_t)$) を計算する。同様に、Actor は現在の環境の状態 (s_t) と Critic で計算された予測報酬を利用して最適方略を学習する。

Actor-Critic 手法で重要な考え方が Critic における報酬の予測誤差である TD 誤差 (δ_t) を Critic での学習だけでなく、Actor の学習にも利用することである。実際には、Critic では報酬の予測がより正確になるように、すなわち、δ_t をゼロに近づけるように学習する。一方で、Actor では環境から得られる報酬を大きくするように、すなわち、δ_t の量に従って行動優先度を増減させる。

Actor-Critic 手法は、行動選択に最小限の計算量しか必要としない点や確率的な行動方策を陽に学習することが可能なことから、強化学習の実装手段として広く利用されている。

2.2 価値関数と予測報酬の最大化

ある方策 π のもとでの状態 (s_t) の価値 (V^π) は次のように定義される。

$$V^\pi(s_t) \quad = \quad E^\pi\{\Sigma_{k=0}^{\infty}\gamma^k r_{t+k+1}\} \tag{1}$$

ここで、E^π はエージェントが方策 π に従ったときの期待値を表し、γ は未来において得られるであろう報酬ほど割り引いて考えるための割引率を表す。γ は 0 から 1.0 の値をとり、$\gamma = 0$ の時は，目前の強化信号にのみ注目し未来を無視することになり、逆に、$\gamma = 1.0$ の時は、行動の評価を遠い未来まで考えることになる。

行動戦略 $\pi(s, a)$ は次式のように決められる。

$$\pi_t(s, a) = \Pr\{a_t = a | s_t = s\} = \frac{\exp(p(s, a))}{\sum_b \exp(p(s, b))} \tag{2}$$

ここで、$p(s, a)$ は時刻 t において、状態 s のとき行動 a を行うことが望ましいかど うかを示す値である。

$p(s_t, a_t)$ は次式により δ_t を用いて修正する。ここで、β は学習率を表す正の定数である。

$$p(s_t, a_t) \quad \leftarrow \quad p(s_t, a_t) + \beta\delta_t \tag{3}$$

2.3 TD 学習

式 (1) を時刻 $t + 1$ について求め、$V(s_t)$ と $V(s_{t+1})$ の差を取ることで、両時刻における状態の評価の間には次の関係が

連絡先: 岡田　浩之，東海大学理学部，〒 259-1292 神奈川県平塚市 北金目 1117，okada@ss.u-tokai.ac.jp

ある。、

$$V(s_t) = E\{r_{t+1} + \gamma V(s_{t+1})\} \quad (4)$$
$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (5)$$

δ_t は、TD 誤差と呼ばれ Critic における報酬予測の精度を表す。

3. 報酬・嫌悪の二次元評価強化学習

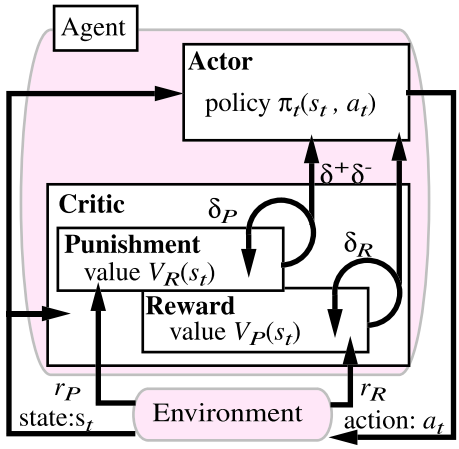

図 2: 二次元評価強化学習

強化学習は本来、報酬を基にした学習の枠組みであるにも関わらず、報酬の扱いがある意味、”雑 ”であり、従来の強化学習研究では正の報酬と負の報酬を単純に足した値を ”報酬 ”として扱っている。ラットやサルでのオペランド条件付け課題や脳に障害を受けたことにより学習機能が損なわれたヒトからの知見から、行動学習に大きな影響を与えるのは成功したときと、失敗したときの評価を区別して考えることだとの指摘があり、我々は、評価関数を報酬性評価と嫌悪性評価の二次元で扱う、報酬・嫌悪の二次元評価強化学習を提案した [1][2] 。

3.1 基本的な考え方

図 2 に提案する二次元評価強化学習の構成を示す。Critic は報酬性評価部 (Reward) および嫌悪性評価部 (Punishment) からなり、それぞれ、環境 (Environment) から状態 (s_t) と報酬性評価 (r_R)、嫌悪性評価 (r_P) を受け取り、それぞれの予測値を学習する。ここで、r_R および r_P は共に正の値を取り、報酬性評価および嫌悪性評価の予測に関する TD 誤差 (δ_R, δ_P) から *Interest*(δ^+) および *Utility*(δ^-) を以下の式で定義する。

$$\delta^- = \delta_R - \delta_P \quad : \textit{Utility} \quad (6)$$
$$\delta^+ = |\delta_R| + |\delta_P| \quad : \textit{Interest} \quad (7)$$

3.11 *Interest* による探索と実行の資源配分

従来の強化学習では報酬性強化信号と嫌悪性強化信号の差 (*Utility*) だけを行動決定の評価として利用している。それに対して、提案する手法では報酬性強化信号と嫌悪性強化信号の和に相当する量 (*Interest*) を定義し、一種の重要度と見なす。重要度は生物における好奇心や動機などと考えることができ、どの処理に注意を向けるべきかを決定することに利用できる。即ち、強化学習に限らず試行錯誤的な学習法では環境同定のための探索と報酬獲得行動の比率を決定することに利用できる。

3.12 予測報酬の時間的割引率の区別

多くの現実的な問題において、報酬性強化信号はいかにしてゴールに辿り着くかに関係し、予測報酬を用いることでゴールに到達するための行動系列を獲得する。従って、遠くにあるゴールからの影響を考慮するために予測報酬の時間的割引率 γ を大きな値に設定する必要がある。一方，危険回避に対応する嫌悪性強化信号に対してはあまりに遠くの状態までその影響を及ぼすと、その効果で多くの入力状態に対して回避行動が発生し、動作主体が探索できる範囲を自ら狭め、結果として動作主体の能力を低下させる場合がある。そのため嫌悪性強化信号には γ を小さくすることで、直前の障害物だけを避ける効果的な行動を生成できる可能性がある。

4. 認知モデルとしての強化学習の可能性

我々が提案する *Interest* は生物における好奇心や動機などと考えることができ、*Interest* を導入した強化学習はヒトの意思決定過程をモデル化する重要な手段の一つになると考えられる。

例えば，失敗した時に感じる、後悔や失望といった感情は人間の意志決定過程を考えるとき、非常に重要である。後悔も失敗も自分が選んだ行動が失敗した際に感じる感情という点では同じだが、後悔は行動を選択する時に候補として考えていた他の行動との比較によって生じるものであり、“ あの時、右に曲がっていればもっと早く到着したのに（実際には左に曲がった）”といったような感情が生じる。それに対し、失望は行動を選択する基準として考えた報酬の予測と実際に得られた結果との比較で生じる。“ 30 分で到着するはずだったのに、1 時間もかかってしまった ”などのように予測と実際の結果の差が大きいほど失望の念は大きい。

5. 終わりに

環境との相互作用を重視する点で起源を同じくする、行動分析学におけるスキナーの一連の研究と Samuel の Checker Player に端を発する強化学習との接点は従来研究では希薄だった。改めて、両研究の接点を見直すことで，教育や福祉、ビジネスといった広範な領域にその研究範囲を拡大できると考える。

最後に、強化学習の ”強化 ”という言葉だけから行動主義との連想で、単純な刺激−反応現象のモデル化手法のように捉えられることがあるが、言うまでも無く、強化学習においてエージェントが環境の内部モデルを持つことは大きな特徴の一つであり、批判の対象となっている行動主義とは違うことを強調したい。

参考文献

[1] H.Okada, H.Yamakawa, and T.Omori.:Neural Network model for the preservation behavior of frontal lobe injured patients. ,ICONIP’98, Vol.3, pp.1465-1469 (1998).

[2] 岡田浩之, 山川宏, 大森隆司.:環境同定と報酬獲得のトレードオフを解消する報酬・嫌悪の二次元評価強化学習の提案, 日本ロボット学会誌, Vol.19, No.2, pp.244-251 (2001).