強化学習

浅川伸一 <asakawa@twcu.ac.jp>

1　はじめに

ニューラルネットワーク研究あるいは機械学習の分野で長いこと手が付けられずに残っていた重要な考え方 —そして，この受業でも取り上げていない問題— があります．それは「生体が何かを欲求し，環境から来るある特定の信号を最大化させるように，自らの行動を適合させる」という，いわば当たり前の考え方，しかし生物の進化や適応にとって重要なメカニズムです．

この考え方は近年「強化学習 reinforcement learning 」という名で注目を集めるようになって来ました．心理学の世界では「道具的条件づけ instrumental conditioning」と呼ばれる学習理論の一般化，あるいはニューラルネットワーク的実装と言える側面も持っていますが，これから紹介する強化学習の枠組みは，行動主義心理学者たちの考える道具的条件付けよりも広くて一般的な意味で用いられます．

たとえば幼児は感覚系 (眼や耳や皮膚) と運動系 (声や手足) との連係を用いて環境 (母親や自分の身の回りのもの) に直接的に働きかけます．この連係を用いることで原因と結果の推論や，目標を達成するために何をすべきかについて多くの情報を取りだすことができます．このような環境との相互作用が我々自身に関する主要な知識源であると言えると思います．アーサー C クラーク「2001 年宇宙の旅」風に言えばおよそ 300 万年前，ヒトザルがモノリスに接触することで,「道具」の使用を発見し，自分の発明した道具を使用した結果と環境との相互作用を学習することによって，この惑星の頂点に君臨し，今日の文明を築いて来たとも言える訳です．

2　強化学習の特徴

強化学習の特徴を挙げるとすれば，試行錯誤的な探索 trail–and–error search と遅延報酬 delayed reward の 2 点になります．行動は直接的な報酬のみならず，その次の情況に影響を与え，そのことを通じて，その後に続く全ての報酬に影響を与えます．いままで扱って来たニューラルネットワークにおける教師あり学習は外界から与えられる教師信号によって自らの行動を適応させていく，例からの学習です．この例からの学習も重要な要素の 1 つですが，

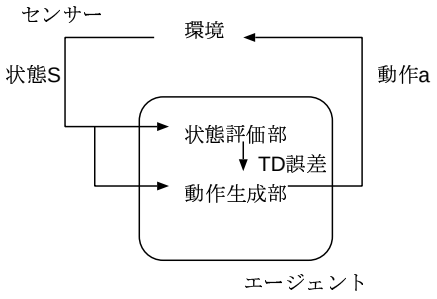
相互作用を介した学習では未知の問題領域で学習者自らが経験から学ぶ必要があるのです.

教師あり学習では，与えられた入力に対して最初はランダムな結合係数によって答えを出し，その答えを教師信号の示す方向に変化させていくと言う意味では結合係数によって定義される空間の探索を行なっていると見なすことができます．一方，強化学習では自身の取りうる行動のレパートリーの中から最適な行動を探索していると見なすことができます.

強化学習には，外の手法と異なり抽象的な概念 —ポリシー，プランニング，価値関数，報酬関数，環境のモデルなど— を直接取り扱います.

ポリシーはある時点での学習者の振舞い方を定義する確率として扱われます．報酬関数は目標を定義します．強化学習者 (エージェント) の目的は最終的に受け取る総報酬を最大化することです．価値関数は最終的に何がよいのかを指定します．ある状態の価値とはエージェントがその状態を起点として将来にわたって蓄積することを期待する報酬の総量です．人間にたとえれば報酬は喜びや苦痛のようなものですが，価値は我々の環境が特定の状態にあるときどれだけ満足あるいは不満であるかに関して，もっと洗練された長期的観点からの判断に相当します．すなわちエージェントはもっとも高い報酬ではなくもっとも高い価値 (しばしば総報酬量の関数として定義される) を持つ状態につながるような行動を見つけ出そうとするわけです.

強化学習が他のタイプの学習ともっとも異なる特徴は正しい行動を直接与えて教示するのではなく，実行した行動の評価を訓練情報として利用することです．従って，よい行動を直接探索するために試行錯誤による能動的な探索が必要になります．行なった行動がどれくらい良いのかが知らされ，それが可能な行動の中で最良または最悪であるかについては知らされません.



それぞれの行動に対して，その行動が選ばれた場合の報酬の期待値が定まっていて，この値を価値と呼びます．強化学習では価値が確実に知られているわけではないと仮定します．その場合でも価値の推定値を持つことが出来ます．行動の価値の推定値を常に持っていれば，どの時点でも，価値の推定値を最大とするような行動がすくなくとも 1 つ見つかることになります．価値の推定値を最大とするような行動を選択することをグリーディ greedy な行

動と呼びます．グリーディでない行動は探索を行なっていると言います．なぜならグリーディでない行動を選択すれば，その価値の推定値を改良できる可能性があるためです．探索はより大きい総報酬を最終的に作りだす可能性を持っています．たまに小さい確率 ϵ でグリーディな行動選択とは無関係に一様に任意の行動を選ぶような方法を ϵ グリーディ法と言います．ϵ グリーディ法ではすべての行動 a に対して行動 a の価値の推定量 $Q_t(a)$ が真の推定量 $Q^*(a)$ に収束することが保証されています．ϵ グリーディ法の欠点の１つは探索を行なう際にすべての行動を等しく選択してしまうことです．つまりほとんど最悪と思われる行動を選択する可能性とほとんど最適行動に近いような良い行動を選択する可能性が同程度に高いことを意味します．これを解決するのがソフトマックス行動基準

$$\frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}} \tag{1}$$

です．ここで τ は温度と呼ばれ温度が高い程全ての行動がほぼ同程度に起こることになります．$\tau \to 0$ の極限ではグリーディ行動選択と一致します．

3　目標と報酬

強化学習ではエージェントの目的あるいは目標は，環境からエージェントに送られる特殊な信号として形式化することができます．各時間ステップにおいて，報酬は単純に数値 $r_t \in \mathcal{R}$ です．非型式的には，エージェントの目標は自分が受け取る報酬の総量を最大化することです．これは，直接的な報酬を最大化することではなく，最終的な累積報酬を最大化することを意味しています．

目標に関する考え方を形式化するために報酬信号を用いることは，強化学習の大きな特徴の１つです．例えばロボットに迷路から抜け出すことを学習させる際には，脱出して報酬が 1 になるときまでは報酬を 0 にすることがよく行なわれます．迷路の学習においてよく行なわれることは，脱出する前のあらゆる時間ステップで-1 の報酬を与えることです．これによってロボットは出来る限り迅速に迷路を脱出するように仕向けられます．たとえば有名なロボット三原則において

ロボット工学の三原則

第一条　ロボットは人間に危害を加えてはならない．また，その危険を看過することによって，人間に危害をおよぼしてはならない．

第二条　ロボットは人間にあたえられた命令に服従しなければならない．ただし，与えられた命令が，第一条に反する場合は，この限りではない．

第三条　ロボットは，前掲第一条および第二条に反するおそれのないかぎり，自己を守らなければならない．

— ロボット工学ハンドブック，第 56 版，西暦 2058 年

第一条の人間に危害を加えることは $-\infty$ の報酬を意味し，第二条の人間の命令に服従することは報酬 ∞ などとなります．すなわちロボットは常に自身が得る報酬を最大化しようとします．もし人間のためにロボットに何かをさせようとするなら，報酬を最大化することでロボットが我々の目標を達成してくれるように報酬を与える必要があります．我々が設定する報酬が我々が達成したいことを真に示していることが決定的要因なのです．もし人類が宇宙に進出し銀河帝国を築くとしたら，そしてその銀河帝国が滅びゆく運命だったとしましょう．このときロボットに与える報酬はなるべく速やかに人類の繁栄を取りもどさせるために，言い替えれば，出来る限り迅速に人類に暗黒時代を切りぬけさせるために報酬が与えられるべきでしょう．別の例で説明するのなら木星ミッションを達成するための報酬とディスカバリー号の乗組員を殺してしまう負の報酬との間で前者に対する報酬が強くしてあったことが 2001 年宇宙の旅で HAL が犯してしまった誤りなのです．

4　時間差分学習 Temporal Difference Learning 法

強化学習の難しさの１つは，必ずしも動作の直後に強化信号が得られないというところにあります．強化学習では数値化された報酬信号を最大にするために，何をすべきかを (どのようにして情況に基づく動作選択を行なうか) を学習します．通常のニューラルネットワークの学習方式のように学習者がどの行動をとるべきかは教えられず，その代わり，どの行動をとればよりいっそうの報酬に結びつくかを見つけ出す必要があります．

時間差分学習 Temporal Difference Learning (TD) 法では時刻 $t+1$ で目標値を作り，観測した報酬 r_{t+1} と価値の推定量 $V(S_{t+1})$ とを使って適切な更新を行ないます．もっとも単純な TD 法は TD(0) と呼ばれ以下のようになります．

$$V(S_t) \leftarrow V(S_t) + \alpha\left[r_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right] \tag{2}$$

$V(S_t)$ は時刻 t における状態 S の価値であり，r_{t+1} とは時刻 $t+1$ すなわち次の時刻における報酬です．γ は割引率，α はステップサイズパラメータと呼ばれます．$0 \leq \gamma \leq 1$ であり，γ が小さいと将来における価値の推定量が低く見積もられることを示しています．逆に γ が 1 に近いと遠い将来に得られるであろう報酬を考慮した評価になります．次の時刻における報酬と次の時刻における価値の推定量が大きければ大きい程ステップサイズパラメータに

比例してその価値の推定量が大きくなります．いかなるポリシー π に対してもステップサイズパラメータが十分小さい定数ならばポリシー π に従うときの価値 (期待収益) V^π に収束する TD アルゴリズムは次のようになります．

- $V(S)$ すなわち状態 S における価値の推定量を初期化し，ポリシー π を評価対象のポリシーに初期化する
- 各エピソード (試行) に対して繰り返し：
  - 状態 S を初期化する
  - 状態 S のとき行動 a (π で与えられる) を決める
  - 行動 a を取り，報酬 r と次の状態 S' を観測する
  - $V(S) \leftarrow V(S) + \alpha\,[r + \gamma V(S') - V(S)]$
  - $S \leftarrow S'$
- S が終端記号なら繰り返しを終了