「従来の強化学習法の考察」

では、従来の強化学習について考察をする。前回の囲碁の学習アルゴリズムにモンテカルロ法（MonteCarlo method,MC）という手法を使っていたが、これはシミュレーションや数値計算を乱数を用いて行う手法の総称である。数値解析では、確率を近似的に求める手法として使われる。機械学習の分野ではモンテカルロ法は強化学習の一種として定義されている。

つまり、状態sから得られる報酬の合計を予測し、それをもとに状態の価値と次に行う行動を決定する。状態価値$V(s)$、行動価値$Q(s,a)\,|\,a=$行動、とすると、

$V(s) \leftarrow V(s) + \alpha[R_t - V(s)]$　　ここでαは学習率$(0 < \alpha < 1)$である。

$Q(s,a) \leftarrow Q(s,a) + \alpha[R_t - Q(s,a)]$

このR_tは、シミュレーションによって得られる報酬の総和を未来に得られるものである。

また、上記強化学習（Reinforcement Learning）とは、ある環境内にあるエージェントが現在の状態を観測し、取るべき行動を決定する機械学習の一種である。エージェントは行動を選択することで環境からより多く得られる合理性の方策（policy）を学習する。学習すべき入力データと出力データのペアが与えられない、すなわち「教師なし学習」である。未知の学習領域を開拓していく行動と既知の学習領域を利用して行動とをバランス良く選択することができるという特徴がある。これには代表的な手法としてTD学習やQ学習がある。また、この環境とは、有限状態のマルコフ決定過程として定式化される。

神経科学においては、黒質緻密部（中脳の神経核）のドーパミン作動性ニューロンから電気記録をとり、その位相性の発火が報酬予測誤差信号をコード化していることが判明し、哺乳類の脳において大脳基底核（大脳皮質と視床、脳幹を結びつけている神経核の集まり）はドーパミンを介して強化学習を行う神経回路としてあるが、まさに上記学習法は脳の学習そのものである。

Q学習（Q‐learning）を簡単に説明しよう。Q学習は、有限マルコフ決定過程において、全ての状態が十分にサンプリングできるようなエピソードを無限回試行した場合、最適な評価値に収束することが理論的に証明されている。実際の問題に対してこの条件を満たすことは困難であるが、この証明はQ学習の有効性を示す要素の一つとして挙げられる。実行するルールに対し、そのルールの有効性を示すQ値という値を持たせ、エージェントが行動するたびにその値を更新する。ここでいうルールとは、ある状態とその状態下においてエージェントが可能な行動を「対」にしたものである。例えばエージェントの現在の状態をs_tとし、この状態で可能な行動がa,b,c,dの4通りあるとする。この時エージェントは4つのQ値、$Q(s_t,a),Q(s_t,b),Q(s_t,c),Q(s_t,d)$を元に行う行動を決定する。行動の決定方法は、理論上では無限回数試行するならランダムでもQ値の収束は証明されているが、現実には収束を早めるため、なるべくQ値の大きな行動が高確率で選ばれるように行う。選択方法としてはある小さな確率εでランダムに選択し、それ以外ではQ値の最大の行動を選択するε-グリーディ手法や遺伝的アルゴリズムで使用されているルーレット選択、以下のようなボルツマン分布を利用したソフトマックス手法などが使用されている。

$$\pi(s,a)=\frac{\exp(Q(s,a)/T)}{\sum_{p\in A}\exp(Q(s,p)/T)}$$

ここで T は正の定数、A は状態 s でエージェントが可能な行動の集合である。

行動を決定した場合、次にその状態と行動の Q 値を更新する。

例として状態s_tのエージェントが行動aを選び、状態がs_{t+1}に遷移したとする。

このとき$Q(s_t,a)$を次の式で更新する。

$$Q(s_t,a)\leftarrow Q(s_t,a)+\alpha\left[r_{t+1}+\gamma\max_p Q(s_{t+1},p)-Q(s_t,a)\right]$$

ここでαは学習率といい後述する条件を満たす数値であり、γは割引率といい、0 から 1 迄の定数である。またr_{t+1}はエージェントがs_{t+1}に遷移したときに得た報酬である。

上記の更新式は現在の状態から次の状態へ移ったとき、その Q 値を次の状態で最も Q 値の高い状態の値に近づけることを意味している。このことにより、ある状態で高い報酬を得た場合は、その状態に到達することが可能な状態にもその報酬が更新毎に伝播することになる。これにより最適な状態遷移の学習が行われる。Q 学習は学習率αが以下の条件を満たすとき、全ての Q 値は確率 1 で最適な値に収束することが証明されている。

$$\sum_{t=0}^{\infty}\alpha(t)\rightarrow\infty$$

$$\sum_{t=0}^{\infty}\alpha(t)^2<\infty$$

この性質のため、Q 学習に関する多くの研究が試されているが、いくつかの問題点も指摘されている。例えば、Q 学習による理論的保証は値の収束性のみであり、収束途中の値には具体的な合理性が認められないため、学習途中の結果を近似解として用い難い。パラメータの変化に敏感であり、その調整に多くの手間が必要であるなどがある。

これらの欠点をどのように克服して最適な学習法を得るか…という研究になると、単なる脳科学の模倣では無理であることが明白である。何故なら、人間が完璧な思考をする生物ではないからである。（第 4 版）

⇒ *cTag*＞意味位相空間ページへ