

失敗と成功の競合における Q-学習の研究

旭川高専 ○谷 昌彦 渡辺 美知子 古川 正志

要 旨

従来の Q-学習は成功報酬のみで学習を行っているため、探索空間が非常に広く学習の収束に時間がかかっている。本研究では、失敗の Q-学習を導入し成功と失敗の学習値を2つのエージェントのように競合させ早い学習を行う Q-学習を提案する。また、数値計算実験により本理論が有効である事を検証する。

1. はじめに

現在 TD 学習や Q 学習等に代表される様々な強化学習が提案されている。しかし、例えば TD 学習では、価値関数の伝搬が1回の試行に1ステップしか進まないため、学習の収束に非常にたくさんの時間が必要となる。また、Q-学習では、探索空間が非常に広い場合立ち上がり時の学習が遅いと言う欠陥がある。このような問題は多くの強化学習においてとも言える事である。本研究では、より早い学習を目指して成功と失敗の競合を用いた Q-学習を提案し、数値計算実験を行う。

2. Q-学習のアルゴリズム

Q-学習は Watkins によって提案された強化学習法である。状態と行動の対に対する評価値 (Q 値) を用いて行動を選択し、その結果によって Q 値を更新して行く学習アルゴリズムである。

ある時刻 t において状態を st 、その時の行動を at とすると Q 値は $Qt(st, at)$ で表される。確率的に行動を決定するために行動決定をボルツマン分布によって行う。

$$P(st, at) = \frac{\exp(Q_t(st, at)/T)}{\sum_{b \in A} \exp(Q_t(st, b)/T)} \quad (1)$$

ここで A は行動集合、 T は温度係数を表す。温度係数 T が高くなると Q 値の差が反映されにくくなる。行動 at を実行した結果、時刻 $t+1$ の時、状態 $st+1$ に移行したとすると $Qt+1(st, at)$ は次式によって更新される。

$$Qt+1(st, at) = (1-\alpha)Qt(st, at) + \alpha[c + \gamma \max_{b \in A} Qt(st+1, b)] \quad (2)$$

ここで α は学習率で、この値が高いと学習速度が速くなる。 γ は割引率で、この値が高くなると長期的な報酬しか考えなくなる。 c は環境からの評価を表す。

3. 問題の設定

図1のようなコース内で車をスタートからゴール

まで移動する事を目的とする。

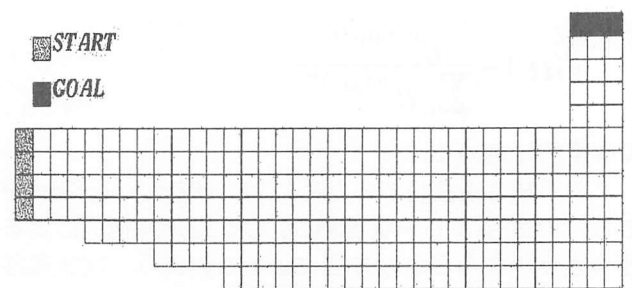


Fig.1 A race track course.

状態は車の位置 (X, Y) 、速度 (Vx, Vy) の4つの値によって決まるものとする。車は1 STEP 毎に加速度 $(Ux, Uy) = \{(-1, -1) (-1, 0) (-1, 1) (0, -1) (0, 0) (0, 1) (1, -1) (1, 0) (1, 1)\}$ の9つから一つを選択し次式により次の状態に移る。

$$\begin{aligned} X_{t+1} &= X_t + V_{xt} + U_x \\ Y_{t+1} &= Y_t + V_{yt} + U_y \\ V_{xt+1} &= V_{xt} + U_x \\ V_{yt+1} &= V_{yt} + U_y \end{aligned}$$

ただし、壁にぶつかった時は速度 $(Vx, Vy) = (0, 0)$ とし $(X_{t+1}, Y_{t+1}) = (X_t, Y_t)$ とする。また、車の初期状態は速度 $(Vx, Vy) = (0, 0)$ とし、スタート位置は4つの中からランダムで選ばれるものとする。

4. 成功と失敗を使用した Q-学習のアルゴリズム

行動は成功と失敗の2つの評価値 (成功の方を Qs 、失敗の方を Qf とする) の比較、選択により決まるものとする。2つの評価値の更新の式は式 (2) であり、ここで環境からの成功の直接評価 Cs 、失敗の直接評価 Cf を

$$\begin{aligned} \text{ゴールしたとき:} & \quad Cs = 10 \quad Cf = 0 \\ \text{壁にぶつかったとき:} & \quad Cs = 0 \quad Cf = 10 \end{aligned}$$

としこれ以外は $(Cs, Cf) = (0, 0)$ とする。

比較、選択の方法として次の3つを提案する。

$$4.1 \text{ 成功 } (Qs) \text{ と失敗 } (Qf) \text{ の差を用いる方法} \\ Q(s, a) = W1 * Qs(s, a) - W2 * Qf(s, a) \quad (3)$$

Q (s, a)を式 (1) に代入して行動を決定する。

但し、W 1、W 2 は Qs、Qf の重みである。

4.2 成功 (Qs) と失敗 (Qf) のボルツマン分布を比較する方法

Qs、Qf をに代入して $P_s(s,a)=P(s,a,Q_s), P_f(s,a)=P(s,a,Q_f)$ を得るものとする。しかし、失敗の分布を逆分布にするため

$$Q(s,a)=\max_{b \in A}(Q_f(s,b))-Q_f(s,a) \quad (4)$$

とする。

行動を決定は $P(s,a)=W_1*P_s(s,a)+W_2*P_f(s,a)$ としこれを正規化したものを用いる。

4.3 成功 (Qs) と失敗 (Qf) の選択による方法 成功の評価値と失敗の評価値の逆分布を比較し大きい方を用いる。

$$Q_f(s,a)=\max_{b \in A}(Q_f(s,b))-Q_f(s,a)$$

$$Q(s,a)=\max(Q_s(s,a), Q_f(s,a)) \quad (5)$$

とし Q (s, a)を式 (1) に代入して行動を決定する。

5. 数値計算実験

α を0.3 γ を0.8とし4章で提案した3つのアルゴリズムについて数値計算実験を行う。それぞれのアルゴリズムで一番良い結果が得られたものを世代とゴールまでにかかったSTEP数のグラフとして4.1、4.2、4.3をそれぞれ図2~4に示す。ここで、アルゴリズム4.1では $W_1=1, W_2=1$ であり4.2では $W_1=5, W_2=1$ である。また、従来のQ-学習と比較するため成功の評価値のみの結果を同図に示す。

6. おわりに

数値計算実験からアルゴリズム4.1、4.3については重みなどの変数を慎重に選択する必要があるが本研究の目的である早い学習を行わせる事が出来た。

アルゴリズム4.2については良い結果を得ることは出来なかった。また、今後はより複雑なコースでの検討などを行う必要があると思われる。

参考文献

- 1) 銅谷 賢治; 強化学習、日本神経回路学第7回全国大会講演論文集、p 158-162
- 2) Andrew G.Barto, Steven J.Bradtko, Satinder P.Singh; Learning to act using real-time dynamic programming, Artificial Intelligence 72, p 81-138
- 3) Richard S.Sutton; Learning to Predict by the Methods of Temporal Differences, Machine Learning 3 p 9-44

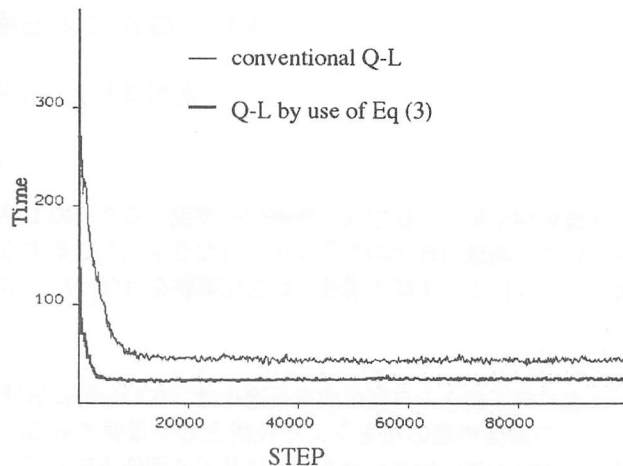


Fig. 2 Comparison of Q-L based on Eq.(3) with the conventional Q-L on learning curves.

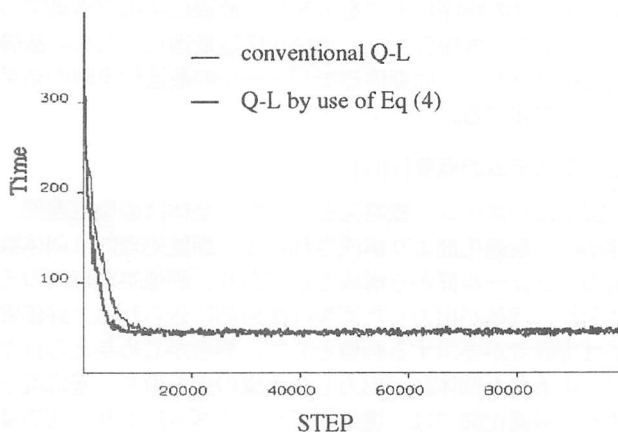


Fig. 3 Comparison of Q-L based on Eq.(4) with the conventional Q-L on learning curves.

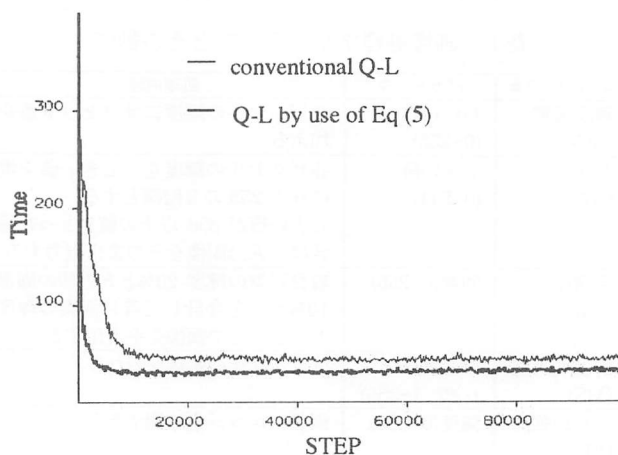


Fig. 4 Comparison of Q-L based on Eq.(5) with the conventional Q-L on learning curves.