メモリQ学習
― 強化学習による選択的記憶の獲得 ―

大分大学　○柴田克成　宮本沢巳
shibata@cc.oita-u.ac.jp

Memory-Q Learning
- Acquisition of Memory Selection by Reinforcement Learning -

Katsunari Shibata and Sawami Miyamoto, Oita University

Abstract: As an approach of reinforcement learning in POMDP (Partially Observable Markov Decision Process), a novel memory-based learning named "Memory-Q Learning" is proposed. The input of an agent is consisted of a present observation and a shift-register-like short-term memory. The agent decides whether the present observation is memorized or not in its memory. The memorization is considered as an action, and Q-value for memorization is assigned and learned as well as the regular Q-learning. In a simulation, only necessary observations were memorized, and the advantage over the other memory-based approaches was confirmed.

１．まえがき

強化学習は、一般的に、「行動」のための学習と捉えられてきている。しかしながら、筆者らは、認識、記憶を始めとするセンサからモータまでの他の一連の機能も、目的を達成するための行動の一種と考えることで、強化学習によって自律的、適応的に獲得することを提唱し、ロボットや生物の「知能」との関連を議論してきた[1][2]。本研究では、「**強化学習で記憶の学習ができる**」ことを示すことを大きな目的とする。そして、わかりやすく、解析が容易な新しい記憶の学習アルゴリズムを提案する。

Q学習は、収束性が示されていること、わかりやすいということから、最もよく使われる強化学習アルゴリズムの一つである。しかし、MDP(マルコフ決定過程)において最適解へ収束することが示されているものの、POMDP（部分観測マルコフ決定過程）においては、学習がうまくいかない場合があることが知られている。これを解決するための方法がいくつか提案されているが、これらの手法はおおまかに、(1)確率的手法（[3]など）、(2)モデルベース手法（[4]など）、(3)メモリベース手法（[4][5][6]など）、(4)リカレントネットを使ってQ値を計算する手法（[4]など）の４つに分類できる。

これらの中で、(1)の確率的手法では、解は得られるものの、最適解に至らないという問題点がある。(2)のモデルベースの手法ではモデルの設定が難しく、観測値をすべて予測するようなモデルでは、不必要な観測値の予測にメモリが浪費されてしまうなどの問題点がある。これに対し、(3)のメモリーベースの手法は、過去の観測値を記憶しておき、それと現在の観測値からQ値を学習するというアプローチであり、非常にわかりやすい。しかし、観測値をそのまま記憶するため、無駄な面が多く、過去どれだけの観測値を記憶したらよいかを予め決定することも難しい。これに対し、(4)のリカレントネットを用いる方法は、過去の入力（観測値や行動）から必要なもののみを強化学習に基づいて合目的的、適応的に選択して文脈として保持し、出力に反映させることができる上、連続値入力にも対応できるため、上記の中で最も有望であると筆者らは考えている。しかしながら、リカレントネットを用いているため、実用的な学習則がなく、その解析が難しいという問題点が存在する。

本研究では、わかりやすさと解析の容易さに重点を置き、いつ何を記憶したかを容易に確認できるように、(3)のメモリベースのアプローチを取る。

過去の(3)のアプローチを見ると、まず、window-Qアーキテクチャ[4]があげられる。これは、固定長の過去の観測値、行動を全て記憶する方法であるが、過去どのぐらいさかのぼって情報を記憶すればよいかを学習前に予め決定することは困難である。

これに対し、Utile Suffix Memory (USM)[5]などでは、決定木の葉ノードを状態とし、記憶する観測値の長さを可変としている。しかしながら、どの時点

から記憶するかは可変であるが、その時点からの観測値は、不必要なものや、記憶容量から考えて記憶しないほうが良いものであってもすべて記憶することになってしまう。

一方、必要な状態にラベルを貼り、そのラベルを用いて自分の状態を把握するlabeling Q学習[6]が提案されている。この手法は、大幅に記憶量を改善できるが、どこにラベルを貼るかは、観測値の変化などに基づいており、タスク達成の必要性から来るものではない。また、スタート地点が固定されていない場合には適用が難しいと予想される。

本論文で提案するメモリQ学習は、タスク達成のために必要な観測値のみを選択的に記憶することを考える。そして、「記憶」を目的達成のための「行動」の一種と考えることによって、現在の観測値を記憶するかしないかを強化学習によって獲得することを提案し、簡単な迷路問題のシミュレーションで検証する。そして、強化学習で「記憶」の学習ができることを示す。

2. メモリQ学習

2.1 記憶方法

始めに、選択的記憶をどのように行うかを説明する。Fig. 1に、選択的な記憶を行った場合の観測値とメモリの変化の例を示す。ここでは、メモリ容量は2とし、2回分の観測値を記憶できる場合について説明する。行った行動もあわせて記憶するという方法も考えられるが、本論文では、記憶容量節約のため、観測値のみ記憶することとする。

まず、行動については、Fig. 1(a)のように、行動によって観測値が変化し、再び行動しと繰り返していく。記憶に関しては、Fig. 1(b)のように、現在の観測値をメモリに記憶するかしないかをエージェントが決める m を定義し、これが1の時には現在の観測値をシフトレジスタであるメモリに記憶し、その代わり、メモリ中の一番古い観測値はメモリからなくなる。m が0の時は、メモリの中身は変化しない。そして、この m を行動と見立てて、Q学習で行動とともに学習していく。これによって、記憶するかしないかがタスク達成に必要かどうかで判断されるようになると期待される。

2.2 学習方法

メモリQ学習では、2種類のQ値を扱う。一つは、通常のQ学習と同様に、行動のためのQ値(Q)、もう一つは、記憶のためのQ値(Q_{mem})である。また、通常のQ学習では、観測値 x と行動 a のペアに対しそれぞれQ値を割り当てるが、メモリQ学習では、

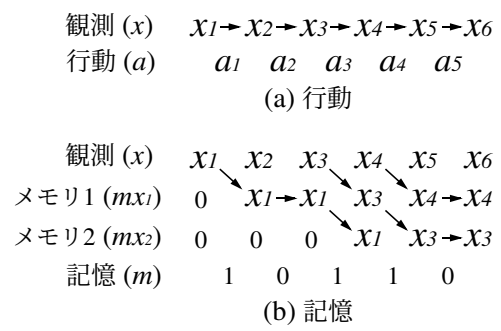

Fig. 1 Change of observation and memories by action and memorization.

行動用のQ値は、観測値 x、メモリに記憶した観測値 mx_i (i=1,..,M; M:メモリ容量)、および行動 a のペアに対して割り当て、記憶用のQ値は、観測値 x、メモリ中の観測値 mx_i、および行動 a の代わりに記憶するかしないかを表す m のペアに対して割り当てる。学習は、メモリ容量が1の場合を例に示すと、

$$\Delta Q(x_t, mx_t, a_t) = \alpha\{r_{t+1} + \gamma \max_{i\in A} Q(x_{t+1}, mx_{t+1}, i) - Q(x_t, mx_t, a_t)\} \quad (1)$$

$$\Delta Q_{mem}(x_t, mx_t, m_t) = \alpha\{r_{t+1} + \gamma \max_{i\in A} Q(x_{t+1}, mx_{t+1}, i) - Q_{mem}(x_t, mx_t, m_t)\} \quad (2)$$

とする。ただし、A: 行動の集合である。(2)式の Q_{mem}の学習で、$\max_{i\in A} Q(x_{t+1}, mx_{t+1}, i)$ の代わりに $\max_{i\in\{0,1\}} Q_{mem}(x_{t+1}, mx_{t+1}, i)$ を用いることも考えられるが、学習がうまくいけば同じ値になると予想されること、わかりやすさから(2)式のようにした。

2.3 行動選択

行動選択は、本論文ではボルツマン選択を用いた。ただし、ゴールから離れてQ値が小さいときにも、その微小な差を行動選択に反映させるために、

$$\overline{Q}(x, mx, a) = \frac{Q(x, mx, a)}{\max_{i\in A} Q(x, mx, i)} \quad (3)$$

と、Q値を正規化し、その値を用いて

$$prob(a) = \frac{\exp(\overline{Q}(x, mx, a)/T)}{\sum_{i\in A} \exp(\overline{Q}(x, mx, i)/T)} \quad (4)$$

とボルツマン選択によって行動の選択確率を計算した。これは、記憶 m の選択も全く同様であり、(3)(4)式の Q を Q_{mem} に置き換えた形である。温度は T は、学習開始時を1.0とし、試行回数とともに指数関数的に徐々に減らし、総試行回数の8割のと

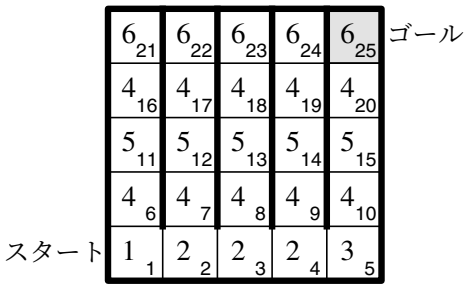

Fig. 2　The maze problem employed in this paper. A large number in each square indicates an observation value, and a small number at lower right indicates a real state.

ころで0.01になるようにし、残りの試行回数では0.01のまま一定とした。

３．シミュレーション

3.1 設定

メモリQ学習の有効性を検証するために、Fig. 2のような迷路を用いてシミュレーションを行った。この迷路では、図中の太線で示してあるように、各列ごとに上からたてに壁が伸びており、エージェントは、自分の上下左右に壁があるかどうかを観察できるとする。また、少し問題を難しくするため、真ん中の行（観測値５）は、その上下の行（観測値４）と区別ができるものとする。エージェントの行動は、上下左右のいずれかへの移動とし、壁にぶつかる行動を選択した場合は、その場にとどまるとした。

この問題では、エージェントが壁にはさまれた縦の列にいる場合、現在の観測値だけでは、自分が５個の列のうち、どこの列にいるかいるかを知ることができない。しかし、一番右の列では、上に、その他の列では下に行かなければゴールに到達しないため、効率よくゴールに向かうためには、同じ観測値の場合でも違う行動を生成しなければならない。また、ゴールから近い状態の観測値４での評価値が増加すると、他の観測値４の場合も評価値が全て同じとなるため、スタート直後に右に行って観測値２を得るよりも、上に行って観測値４を得た方が良いことになり、学習が進まないことが考えられる。

また、観測値を記憶できるメモリを持っている場合でも、常時記憶する方法では、メモリ容量が２以下の場合、上の方に行くと、自分が観測値３(状態５)を通ったかどうかを忘れてしまうことになる。したがって、ゴールに到達するためには、観測値が４や５であっても、それを記憶せず、観測値３の状態を通ったという記憶を保持しておく必要がある。

Q値の初期値はすべて0.5、学習率αは0.1、割引率γは0.9とした。また、ゴールに到達したら、エージェントは再びスタート地点に戻るとした。

比較のため、本論文で提案しているメモリQ学習（メモリ１個）以外に、メモリがない通常のQ学習、および、メモリを一つ設け、常に観測値を記憶する場合の３つについてシミュレーションを行った。

3.2 結果

100回のシミュレーションを行い、総試行回数を変化させたときのスタートからゴールまでの平均ステップ数をFig. 3に示す。ただし、１試行あたりのステップ数は1000で打ち切り、その場合はすべて1000ステップとした。総試行回数が少ない場合をFig. (a)に、総試行回数が多い場合をlogスケールでFig. (b)に示す。メモリQ学習では100試行程度でほぼ最適値（８ステップ）に落ちている。

一方、メモリがない場合と常時記憶する場合は、いずれも、学習前の全くランダムな行動をとった場合の500弱のステップ数から、学習が進むといったん増大するが、数100ステップのあたりから減少している。この時、各シミュレーションでの到達ステップ数を見てみると、メモリがない場合は、ほとんどの場合、最短ステップ数で到達するか、1000ステップ内で到達できなかったかのいずれとなっており、その平均として中間的な値が出ている。一方、常時記憶する場合は、最短ステップでも1000ステップでもない中間的な値を取る場合が多かった。

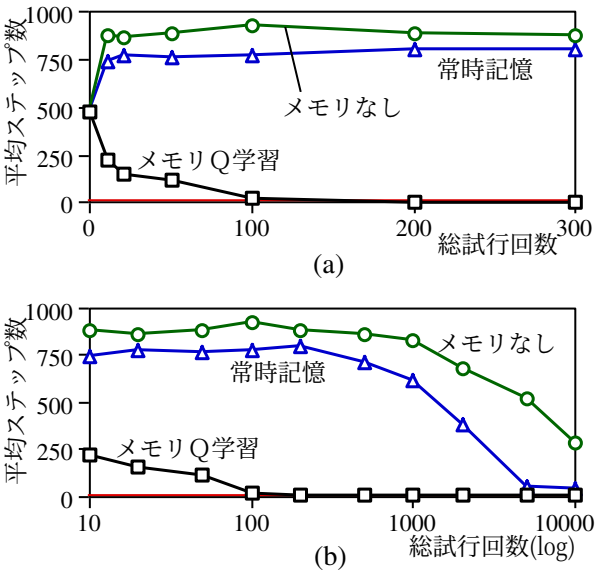

Fig. 3 Necessary steps to the goal according to the total number of trials.　The value is the average over 100 simulations.

最初にステップ数が増大するのは、状態５に行く前にエージェントが上に行き、そこでトラップされてしまい、1000ステップ内でゴールに到達しないことが多くなっているためである。メモリがない場合に、最適解を学習することがある理由としては、試行回数を増やすと、温度がゆっくりと下がるため、たまたま最短もしくはそれに近い経路を通ってその時のＱ値がいったん上昇すると、温度が低い状態では、観測値３(状態５)にたどり着く前に上に行くことがほとんどなくなり、結果的に最短経路を取る場合があると考えられる。また、メモリがあり、常時記憶する場合は、状態２や３や４で上に行き、再び戻ってからゴールに向かうという経路が学習されていることが観察された。

次に、メモリＱ学習で10000試行の学習をした後の主なメモリ用Ｑ値を観察したものをFig. 4に示す。観測値が３(状態５)の場合、記憶するＱ値が大きくなっており、逆に、観測値が４か５で、メモリの中の観測値が３の場合、記憶しない方のＱ値が大きくなっている。このことより、記憶すべきところは記憶し、してはならないところではちゃんと記憶しないように学習できていることがわかる。

一方、現在の観測値が２で、メモリ中に保管されている観測値が３の場合は、記憶しない方のＱ値が大きくなっている。この場合、Fig. 5のように、現在の観測値を記憶すると状態５(観測値３)は忘れてしまうが、記憶しないで、上に行く行動をとれば、状態５の上にいる場合と、現在の観測値、メモリ中の観測値ともに同じになるため、Ｑ値が大きくなる。したがって、状態４において、メモリ中に保管されている観測値が３の場合は、記憶しない方のＱ値が大きくなってしまうと考えられる。このような経路は解の中に含まれず、最適経路が得られたものの、このＱ値は合理的ではないと言える。そもそも、現在のメモリＱ学習のアルゴリズムにおいては、状態の区別をすること自体では報酬が得られないため、より良い状態と混同することが起こる可能性があると考えられる。本件については、今後より深い検討が必要である。

Q_mem(3,2,0) = 0.126
Q_mem(3,2,1) = 0.759
Q_mem(4,3,0) = 0.843
Q_mem(4,3,1) = 0.218
Q_mem(5,3,0) = 0.743
Q_mem(5,3,1) = 0.111
Q_mem(2,3,0) = 0.405
Q_mem(2,3,1) = 0.156

Fig. 4 Some Q-values after learning

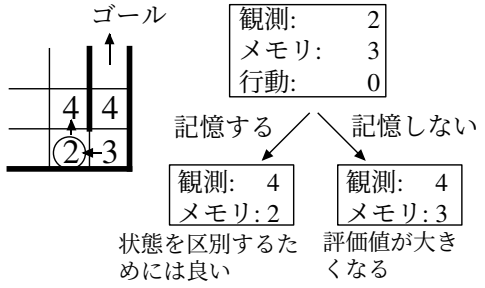

Fig. 5 Problem: Q-value for non-memorization is larger even if the observation should be memorized.

４．まとめ

記憶するかしないかを強化学習によって学習させるメモリＱ学習を提案した。簡単な迷路問題において、必要なところを記憶し、覚えてはいけないところでは記憶をしないで最短経路を獲得することができ、常時記憶する場合と比較して、同じメモリ容量でも、学習がうまく進むことを示した。

謝辞

本研究は、日本学術振興会科学研究補助金基盤研究(B)(14350227, 15400064)の補助の下で行われた。ここに謝意を表します。

参考文献

[1] 柴田克成. "強化学習とロボットの知能 -あめとむちで知能は作れるか？-", 第 16 回人工知能学会全国大会論文集, パネルディスカッション「強化学習とその諸相」パネリスト資料, 2A1-05 (2002)

[2] 柴田克成, 岡部洋一, 伊藤宏司, "ニューラルネットを用いた Direct-Vision-Based 強化学習 － センサからモータまで－ ", 計測自動制御学会論文集, 37 巻 2 号, 168-177 (2001)

[3] Williams, R.J., "Simple Statistical Gradient- Following Algorithm for Connectionist Reinforcement Learning", Machine Learning 8, 229-256 (1992)

[4] Lin, L.J. and Mitchell, T.M., "Reinforcement Learning with Hidden States", Proc. of 2nd Int'l Conf. on Simulation of Adaptive Behavior, 271-280 (1992)

[5] McCallum, R.A., "Instance-Baced Utile Distinctions for Reinforcement Learning with Hidden State", Proc. of the 12th ICML, 387-395 (1995)

[6] Lee, H.Y., Kameya, H. and Abe, K., "Labeling Q-learning for Maze Problem with Partially Observable States", Proc. of ICCAS, 5-8 (2001)