Q学習によるコミュニケーション学習における状態混同の発生

大分大学　○仲西賢展　杉坂政典　柴田克成

Occurrence of state confusion in the learning of communication using Q-learning

Masanobu Nakanishi, Masanori Sugisaka, Katsunari Shibata, Oita University

Abstract: The learning of one-way communication using the Q-learning was verified. A transmitter agent learned what communication signals should be transmitted and a receiver agent learned to generate appropriate actions from the signals. We discovered that when there exists a non-looped branch in the receiver's state transition, and the optimal action in a detour is the same as the optimal one in a state closer to the goal on the optimal path, there is a possibility that the receiver cannot take the optimal path because of state confusion. The reason why the receiver agent falls into the state confusion can be considered that it is not reinforced for the transmitter agent to learn to transmit the value of the state to the receiver agent.

1. まえがき

群ロボットやマルチエージェントシステムにおいて、コミュニケーションは観測の不十分さを補ったり、利害の衝突を避けて協調的な行動を行う上で非常に重要な役割を担う。ロボットやエージェントに目的に沿ったコミュニケーションを自律的・適応的に獲得させるために、進化的手法[1]や強化学習の適用[2][3]が試みられている。文献[1][2]では受け手が行動をする際の観測の不十分さを補うための一方向コミュニケーションを進化的手法や学習によって獲得できる例が示されている。しかしながら、コミュニケーション信号として、どういう情報を伝達すればよいのか、また、あらゆる場合に学習がうまくいくのかどうかといった点については十分な考察がなされていなかった。

筆者らは2エージェント間の一方向コミュニケーションを、強化学習アルゴリズムの一つであるQ学習で獲得させる問題について、受信側がどういうコミュニケーション信号が必要で、発信側がどういうコミュニケーション信号を生成するようになるのかを実験的に検証した。そして、発信側が学習によって生成したコミュニケーション信号によって受信側が状態混同を起こし、受け手が部分観測状態に陥って行動が最適にならない場合があることを発見した。本稿では、その例を示すとともに、どういう時にそのような状態混同が起こるのかをシミュレーションを元に考察した結果を示す。

2. 一方向コミュニケーションの学習

2-1 全体の構成

本稿では、前述のように、2 エージェント間で、片方のエージェントの観測の不十分さをもう片方のエージェントが補うためのコミュニケーションに焦点を当てる。

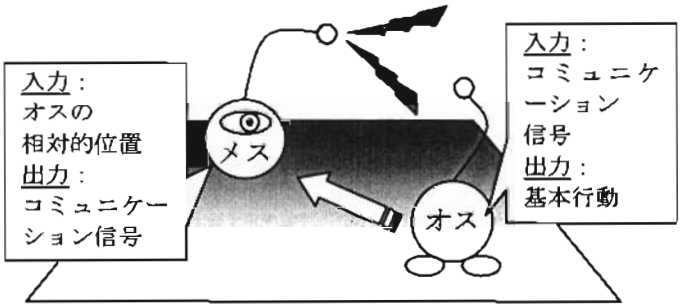

Fig.1 一方向コミュニケーションのイメージ図

そして、文献[1][2]を参考に、オス、メスと呼ばれる 2 体のエージェントを仮定し、両者が接触したら、両者共に報酬がもらえるようなタスクを考える。構成図を Fig.1 に示す。メスは視覚を持ち、オスの位置を特定することができるが、移動することができない。一方、オスは足を持ち移動できるが、視覚を持たず、メスの位置を直接特定することはできない。そして、メスは信号を発生し、オスはメスが発生する信号を正確に識別できるとし、オスの行動による状態遷移はすべて決定論的であるとした。ただし、コミュニケーション信号の意味は全く与えない。したがって、メスはどういう状態のときにどういう信号を送れば良いか、オスは送られてきた信号をどう解釈し、どう行動に反映させれば良いかを学習する。そして、オスとメスの間で何らかの共通の言語を確立することができれば、効率よく接触を繰り返すことが期待できる。これらのエージェントは次のサイクルに従って行動する。

(1) メスはオスの状態を特定し、信号を発生する。
(2) オスは、メスの信号を特定し、基本行動を実行する。
(3) 両者が接触したら試行終了とし、それ以外は時間ステップ t を $t+1$ へ進めて(1)に戻る。

2-2 発信側および受信側のエージェントの学習

ここで、発信側および受信側のエージェント、すなわちオスとメスの学習について示す。両者の学習はQ学習で行う。Q学習は状態と行動の組に対して評価を行い、その評価値に基づいて確率的に行動を選択するとともに、その評価の方法を学習していく方法である。この方法では離散的な環境、行動を前提とすることが多く、この状態と行動の組に対する評価値をQ値と呼ぶ。

Q学習のアルゴリズムを以下に示す。

(1) エージェントは状態 s_t を観測する。
(2) エージェントは任意の行動選択方法にしたがって行動 a_t を実行する。
(3) 環境から報酬 r_{t+1} を受け取る。
(4) 遷移後の状態 s_{t+1} を観測する。
(5) 以下の更新式よりQ値を更新する。

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

ただし α は学習率 $(0<\alpha\le 1)$、γ は割引率 $(0\le\gamma<1)$ である。

(6) 時間ステップ t を $t+1$ に進めて(2)に戻る。

ここでは、メスの状態 s はオスの相対的位置、行動 a はコミュニケーション信号であり、オスの状態 s はコミュニケーション信号で、行動 a は基本行動である。

2-2 行動選択方法

Q学習の学習過程における行動選択方法としては、ボルツマン（Boltzmann）選択を用いる。この方法は状態 x において行動 a を選択する確率 $p(a|x)$ を以下のように定義したものである。

$$p(a|x) = \frac{\exp(Q(x,a)/T)}{\sum_{i\in A}\exp(Q(x,i)/T)} \quad (2)$$

ただし、A は基本行動の集合、T は温度係数である。T は値が大きいほど選択はランダムになり、積極的に探索を行うことになる。逆に T を0に近づけると、わずかなQ値の差が行動選択に大きく影響することになり、極限では、最大のQ値を持つ行動を選択することになる。本稿のシミュレーションでは、温度係数 T は学習開始時に1.0とし、総試行回数の80%のところで0.01になるように、試行回数とともに指数関数的に徐々に小さくした。そして残りの試行では0.01に固定した。

3. シミュレーション

3.1 ケース1

はじめに、学習がうまくいった例として、文献[1][2]を参考にしたFig.2のようなシミュレーション環境で学習した結果を述べる。

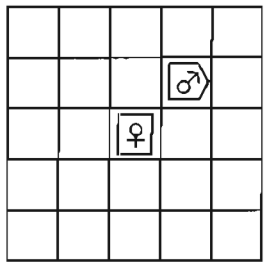

Fig.2 シミュレーション環境1

縦横5×5の離散的な環境（上端下端、左端右端は隣接）で、メスを中心に固定し、オスを各試行毎にランダムに配置する。ここでメスはオスの座標・向きを検出し、0～2の3種類の信号を発生し、オスはメスが発生する信号を受信して基本行動の一つを実行する。ここで基本行動とは、前進(目前のセルへ移動)、右回転(右に90度回転)、左回転(左に90度回転)、静止(状態は変わらない)の4通りである。オスを配置してから、接触するまでを1試行とする。

1000000試行学習を行った結果をFig.3に示す。Fig.3は、各正方形がオスの位置、その中の4つの部分がオスの向きを表している。また、まっすぐな矢印は直進を、曲がった矢印はその方向に回転したことを示す。図中の数字は、例としてオスが右上のセルで右を向いているときの状態遷移の順番を示したものである。Fig.3を見ると、全ての状態(相対距離X座標×相対座標Y座標×オスの向きの96通り)でオスは最適な行動を学習することができた。メスの信号を観察すると、3つの信号が、オスに前進、右回転、左回転の行動を促す信号になり、静止行動は学習を進めていくうちになくなった。つまり、この場合はオスが状態の混同を起こしているにもかかわらず、最適な行動を獲得していることになる。

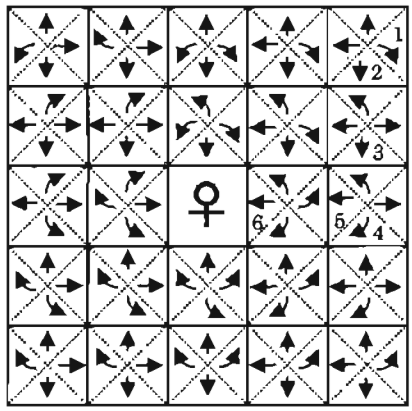

Fig.3 学習後のオスの行動

ただし、信号数を増加させると学習が加速することが確認されている。例えば、総試行回数が 100000 では信号が 3 種類だと最適解を得られることはほとんどないが、信号数を 8 種類に増やすと、10 回中 5 回最適解が得られた。また、試行回数を 1000000 に増やすと、3 種類の信号では 10 回のうち 3 回最適解に収束しなかったが、8 種類だと 10 回中 10 回最適解が得られた。8 種類の時には、一部評価値(ゴールまでの最短ステップ数)による信号の違いも観察され、これが学習の加速につながったのではないかと考えられる。

以上のことから、信号数が多い方が少ない試行で学習ができるものの、オスが最適パスをとるために、行動の数だけ信号があれば、状態評価の情報を信号として送らなくても学習自体は可能ではないかと考えられる。

3.2 ケース 2

次に学習がうまくいかなかった例について述べる。

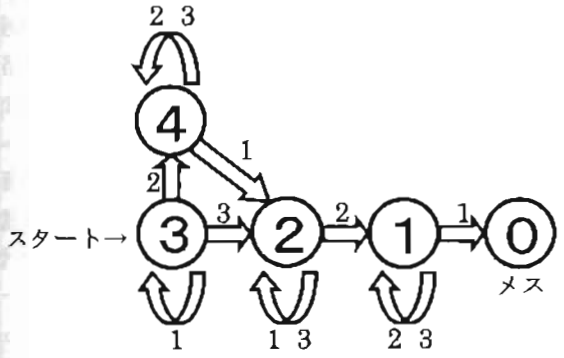

Fig.4 シミュレーション環境 2

Fig.4 にオスの状態の遷移図を示す。丸の中の数字が状態を示し、矢印が状態遷移を示す。オスの行動は 1 ～3 の 3 種類あり、矢印のところの数字はとった行動を表す。ここでメスの入力は現在のオスの状態 0～4 の 5 種類で、メスの発する信号は 1～4 の 4 種類である。また、オスの入力はメスが発する信号で、行動は 1～3 の 3 種類である。よって、信号の種類はオスの行動の種類よりも 1 つ多い。オスの初期位置は状態 3 で、0 がゴール、つまりメスがいる場所である。0 に到達すれば両者に報酬を与え、オスは初期位置に戻る。スタートに配置されてゴールするまでを 1 試行とする。

10000 試行学習を行った後の、オスとメスの Q 値を Table 1 に示す。網掛けで示したのはその状態でもっとも Q 値の高い信号または行動である。

Table 1 を見ると、メスの状態 1,4 では信号 2 を、状態 2,3 では信号 1,3,4 のいずれかを出すようになり、オスは信号 1,3,4 を受信したときは行動 2 を、信号 2 を受信したときは行動 1 を起こすように学習した。このため、状態 3 においても 2 の行動をとってしまい、信号数がオスの行動数よりも多かったにもかかわらず、最適な行動が学習できなかった。メスの信号の数をさらに増やしてシミュレーションを行ったところ、信号が 4 つの場合と同じように、信号の 1 つがオスに行動 1 を促す信号になり、残りすべてがオスに行動 2 を促す信号となったため、やはり最適な行動を学習することがなかった。

Table 1 学習後の Q 値

メスの Q 値 [状態][信号]	オスの Q 値 [信号][行動]
m_q[1][1]=0.610026	o_q[1][1]=0.676914
m_q[1][2]=1.000000	o_q[1][2]=0.754085
m_q[1][3]=0.726789	o_q[1][3]=0.678451
m_q[1][4]=0.670422	o_q[2][1]=0.848748
m_q[2][1]=0.900000	o_q[2][2]=0.698509
m_q[2][2]=0.794080	o_q[2][3]=0.680008
m_q[2][3]=0.900000	o_q[3][1]=0.677502
m_q[2][4]=0.900000	o_q[3][2]=0.756786
m_q[3][1]=0.729000	o_q[3][3]=0.677736
m_q[3][2]=0.652926	o_q[4][1]=0.677074
m_q[3][3]=0.729000	o_q[4][2]=0.757831
m_q[3][4]=0.729000	o_q[4][3]=0.683408
m_q[4][1]=0.725835	
m_q[4][2]=0.810000	
m_q[4][3]=0.728820	
m_q[4][4]=0.727768	

最適な行動が学習できない理由を考察した結果を以下の Fig.5 にまとめる。

状態 1 メスがある信号 A を出したときオスが行動 1 をとりゴールすることを学習し、メスも信号 A を出すことを学習する。

↓

状態 4 メスが信号 A を出すとオスが行動 1 をとるため、メスが信号 A を出すことを学習する。

↓

状態 4 オスは状態 1 の場合と同じ信号を受信するため、状態 4 の評価値が上がる。

↓

状態 3 オスにとって行動 2 の評価が行動 3 の評価より高くなるため行動 2 をとるように学習する。

↓

最適な行動が学習できない

Fig.5 最適行動が獲得できなかった理由の考察

結局、メスは、オスが最適な行動を行うような信号を発信することを学習するために、オスが状態の混同をしてしまうことが問題であることが分かった。信号数をオスの行動数に対して冗長にしたにもかかわらず、コミュニケーション信号において状態 1 と状態 4 を区別できないことは、1 以外の信号の時にオスの行動がすべて 2 になってしまうことが原因の 1 つと考えられる。その理由として考えられることを以下にまとめる。

① 状態 2:メスが信号 A とは別のある信号 B を出した時、オスが行動 2 をとって状態 1 に遷移することを学習する。
② 状態 3:メスはオスに行動 3 をとってほしいため A, B 以外の信号 C を出すように学習する。
③ 状態 3:オスは Fig.5 の状態混同のため、状態 2 より状態 4 の方が評価値が高いため、信号 C に対しても行動 2 をとるように学習する。
④ 状態 3:信号 C の評価が下がり信号 D を出すようになる。

そして、②,③,④をくり返すことにより、オスは A 以外の信号はすべて行動 2 に割り当てるように学習してしまうと考えられる。

3.3 ケース 3

最後に、どのような時に状態混同が致命的になるかを調べるため、すでにオスが行うべき行動とメスが発信する信号が一対一に対応しており、最適経路上でかなりの数の状態混同が起こる場合を考えて学習をさせた。

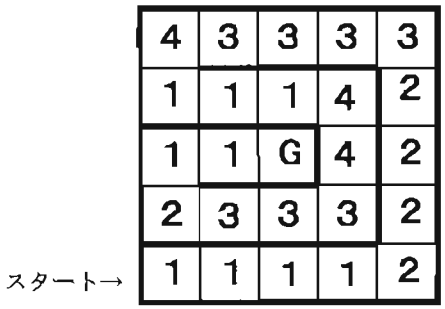

Fig.6 シミュレーション環境 3

ここで、太い線は壁で、左下がスタート、中心がゴールである。オスの行動は上下左右のセルに移動する 4 通りで、壁に向かう行動はその場にとどまるとする。この環境では、Fig.6 のセル内に示した数字は、オスが受け取る信号であるとし、ゴールへの最適行動に対応している。右の場合は信号 1、同様に上は信号 2、左は信号 3、下は信号 4 である。

この環境はオスにとって状態混同のため POMDP(部分観測マルコフ決定過程)となる。しかし、各状態において最適行動以外のすべての行動に対応する Q 値はほぼ 0 となり、正しい行動を学習することができた。したがって、最適経路上で状態混同が起っても、ループ以外の分岐がなければ評価値が下降することがあっても最適経路を通る学習ができると考えられる。以上のことから、状態遷移にループ以外の分岐があり、そのう回路上のある状態での最適行動が、最適経路上のよりゴールに近い状態での最適行動と一致する場合に状態混同が起きて、分岐点においてう回路を通る行動を選択してしまう可能性があることが分かった。逆に、ループ以外の分岐がないとき、メスは最適行動の情報をコミュニケーション信号で表現するようになるため、学習がうまくいくと考えられる。

ケース 1 では、遠回りの経路と状態の混同が起こるにもかかわらず学習はうまくいった。その理由は、状態の混同が起きても分岐点において遠回りの経路の方が評価がよくなることがなかったためと考えられる。

4.まとめ

本稿では、Q 学習により一方向コミュニケーションを学習させる場合に、“状態遷移にループ以外の分岐があり、そのう回路上のある状態での最適行動が、最適経路上のよりゴールに近い状態での最適行動と一致する場合に状態混同が起き、最適経路が得られなくなる”という問題点を明らかにした。さらに、発信者は受信者に最適な行動を伝達するように学習するため、受信者が状態混同を起こすことが原因として考えられるということを示した。また、最適経路上で状態混同が起こっても、ループ以外の分岐がなければ、評価値が下降することがあっても、最適行動を通る学習ができることを示唆した。今後はこの問題点を回避するための手法を考えていく必要がある。

謝辞

本研究は、本学術振興会科学研究費補助金基盤研究(B)(14350227, 15300064)の補助の下で行われました。ここに謝意を表します。

参考文献

[1] G.M. Werner & M.G.Dyer : Evolution of Communication in Artificial Organizing System, Proc.of Artificial life II, 1/47(1991)

[2] N.Ono, T. Ohira, and A. T. Rahmani: Emergent Organization of Interspecies Communication in Q-Learning Artificial Organism,*Advances in Artificial Life*, pp. 396-405 (1995)

[3] 柴田克成, 伊藤宏司:利害衝突回避のためのコミュニケーションの学習-リカレントニューラルネットワークを用いたダイナミックコミュニケーションの学習- 計測自動制御学会論文集 Vol.35, No.11, 1346-1354 (1999)