フェロモンコミュニケーションを行うエージェントシステムにおける移動ルールの獲得

柴　田　淳　子

1. はじめに

多数のエージェントによって構成されたシステムを用いてシミュレーションを行うエージェントシミュレーションが，社会現象を分析するアプローチとして注目されている。これは，エージェントの行動やそれらの相互作用を通じて，システム全体として創発する現象を利用して，現実の社会システムにおける組織形成などの創発現象を解明するものである[1]。

その中の一つである Ant System は，社会性昆虫である蟻が餌を集める行為をモデル化したシステムである。蟻は，餌を集めるために，限定された範囲内の環境を認知し，行動ルールにしたがって移動する。社会性昆虫の特徴は，個々の蟻は単純なルールにしたがっているにもかかわらず，全体として優れた社会的知能を発揮し，より多くの餌を集めることができることである。つまり，集団となることで，より効率的に目標を達成することができる。このような社会性昆虫の特徴を応用した研究は，多くの分野でなされている。車谷は，協調しながら採餌行動を行う蟻コロニーをモデル化し，全体の挙動を表すマクロモデルを生成し，コンピュータシミュレーションにより分析を行っている[2]。また，Dorigo らは，蟻のフェロモンコミュニケーションに基づいた組み合わせ最適化問題の解法アルゴリズムを提案した[3]。このアルゴリズムは，Ant Colony

Optimization (ACO) と呼ばれ，TSP などの問題に適用され有効性が示されている。

一般に，生物は環境の変化に適応するために，自分の行動ルールを変更する。そのために生物は試行錯誤的な行動を繰り返し，その過程で得られた情報をもとに行動ルールを確立する。従来の Ant System において，エージェントの移動ルールは人の手によって事前に設計されている。そこで，本研究では強化学習[4]を用いてエージェントが行動の結果感知したフェロモンの量に基づいた学習を行い，自分自身の移動ルールを獲得する Ant System を構築する。

2. Ant System

Ant System は，エージェントである蟻とその食料である砂糖，そしてエージェントの巣が配置された空間分布である。この空間は，図1のように2次元空間で表現される。蟻の行動は非常に単純であり，餌である砂糖を獲得するために空間上を移動する。そして，砂糖を得た後，フェロモンと呼ばれる化学物質を散布しながら巣に帰る。蟻は，全体としてより多くの砂糖を集めるという目標を認識せずに行動するが，フェロモンを介したコミュニケーションを行うことで，全体として砂糖を効率的に集めることができる。

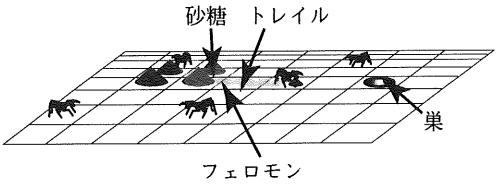

図1　Ant System

エージェントは餌場に到着すると，そこから1単位の砂糖を巣まで持ち帰る。その際，エージェントは餌場から巣までの帰り道に一定量のフェロモンを散布しながら移動する。このフェロモンの道筋をトレイルと呼び，2次元空間の座標 (x, y) に存在するトレイルの濃度 $T(x, y)$ は以下の式によって与えられる。

$$T(x, y) \leftarrow (1-\gamma_{eva})\, T(x, y) - \Delta T_i(x, y) \qquad (1)$$

ただし，

$$\Delta T_i(x, y) = \begin{cases} M_p & (x, y) \text{にエージェント} i \text{がフェロモンを} \\ & \text{分泌した場合} \\ 0 & (x, y) \text{にフェロモンが分泌されなかった場合} \end{cases} \qquad (2)$$

であり，γ_{eva}はトレイルの蒸発定数である。また，座標 (x, y) に存在するエージェントiが，もし砂糖を持っているのであればM_p単位のフェロモンを分泌する。トレイルの蒸発にともない，座標 (x, y) 上のフェロモン濃度$P(x, y)$も変化する。

$$\begin{aligned} P(x, y) \leftarrow & P(x, y) \\ & + \gamma_{dif}\{P(x-1, y) + P(x+1, y) + P(x, y-1) \\ & + P(x, y+1) - 5P(x, y)\} + \gamma_{evl}\, T(x, y) \end{aligned} \qquad (3)$$

ただし，γ_{dif}はフェロモン拡散定数である。

このような環境の中で砂糖を集めるエージェントの行動パターンは，探索，誘引，追跡，運搬の4通りに分けることができ，その詳細は以下の通りである。

(1)探索モード；

初期状態において，すべてのエージェントが「探索モード」である。このとき，エージェントは2次元空間上をランダムに移動する。

(2)誘引モード；

砂糖を獲得していないエージェントが空間上のフェロモンを感知した場合，誘引モードに遷移する。このとき，エージェントはフェロモンの濃度が濃い場所へ移動するという行動ルールに従う。

(3)追跡モード；

探索モードあるいは追跡モードのエージェントがトレイルを感知すると，追跡モードに遷移する。このときエージェントは，トレイルに沿って巣と逆の向きへ移動するという行動ルールに従う。

(4)運搬モード；

エージェントが砂糖を獲得した場合，運搬モードに遷移する。このときエージェントは，砂糖をもったまま巣までの距離を最短距離で運搬する。そして，砂糖をもったエージェントが巣に到着すると，コロニー全体の餌所有量が1単位増加する。

上述のように，個々のエージェントの行動は非常に単純であるが，フェロモンを介した相互作用を通して，Ant System 全体としてより複雑な挙動を創発し，砂糖を効率的に収集することができる。

3. 強化学習

各離散時間ステップ $t=0, 1, 2, \cdots\cdots$ において，エージェントは，現在の環境から状態 $s(t)\in S$ を観測し，行動 $a(t)\in A$ を起こす。その結果，環境は状態 $s(t+1)\in S$ に遷移し，エージェントは報酬 $r(t+1)$ を受け取る。強化学習はこのサイクルの繰り返しによって，具体的にはエージェントとそれを取り巻く環境との相互作用によって進行する。ここで，エージェントの目的は，このようなサイクルの中で，できるだけ多くの報酬を獲得することである。環境の状態は，次の状態遷移関数に従って遷移する。

$$P(s, a, s') = \Pr(s(t+1)=s' \mid s(t)=s, a(t)=a) \qquad (4)$$

ただし，$\Pr(s' \mid s, a)$ はエージェントが環境の状態 s を観測し行動 a を実行した後，環境の状態が s' へ遷移する確率を表す。そして，行動 a を実行した結果，エージェントが環境から受け取る報酬 $r(t+1)$ の期待値は報酬関数として表される。

$$R(s, a) = E\{r(t+1)=r \mid s(t)=s, a(t)=a\} \qquad (5)$$

ただし，$E\{r \mid s, a\}$ は，環境の状態 s においてエージェントが行動 a を起こしたときに受け取る報酬 r の期待値である。このような環境下で，エージェントは観測した環境の状態から最適な行動を導く規則である政策を決定する。

強化学習の代表的な手法としてQ学習がある。Q学習では，Q値と呼ばれる関数 $Q(s, a)$ の値が，実際にエージェントが実行した行動から得られた報酬を

もとに更新され，エピソード数を重ねるにつれて最適な行動を導く政策に近づいていく。ここで，$Q(s, a)$ は状態 s において行動 a を起こす価値を表す関数であり，すべての環境 $s \in S$ と行動 $a \in A$ に対して値が存在する。

Q学習の学習アルゴリズム

Step 1：Q値 $Q(s, a)$ を初期化する。

Step 2：状態を初期状態にもどす。

Step 3：時刻 $t(t=0, 1, \cdots\cdots, t_{\max})$ の環境の状態 $s(t) \in S$ において，エージェントはQ値にもとづいて決定される政策 $\pi(s(t), a)$ にしたがって行動を確率的に選択する。ここでは行動選択法として代表的なボルツマン選択法を採用した。

$$\pi(s(t), a) = \frac{\exp\left(\frac{Q(s(t), a)}{T}\right)}{\sum_{b \in A} \exp\left(\frac{Q(s(t), b)}{T}\right)} \qquad (6)$$

ただし，政策 $\pi(s(t), a)$ は状態 $s(t)$ で行動 a を選択する確率であり，T は温度パラメータである。

Step 4：エージェントは Step 3 で選択した行動 $a(t)$ を実行する。その結果，環境の状態は $s(t)$ から $s(t+1)$ に遷移し，エージェントは環境から報酬 $r(t+1)$ を受け取る。そして，状態 $s(t)$ において行動 $a(t)$ を実行するQ値は次の式に従って更新される。

$$Q(s(t), a(t)) \leftarrow (1-\beta) Q(s(t), a(t)) + \beta \left(r(t+1) + \gamma \max_{a' \in A} Q(s(t+1), a')\right) \qquad (7)$$

ただし，$\beta (0 < \beta \leqq 1)$，$\gamma (0 < \gamma \leqq 1)$ は，それぞれ学習率，割引率を表すパラータである。

Step 5：学習の終了条件を満たす，もしくは最大時刻 $t_{\max}$ であれば学習終了。そうでなければ，時刻を更新して Step 3 に戻る。

Step 6：最大エピソード数であれば終了。そうでなければ，エピソード数を更新して Step 2 に戻る。

4. Q学習に基づくエージェントから成る Ant System

次に，環境に応じた移動ルールを試行錯誤的に獲得するために，Q学習を実装したエージェントから構成される Ant System を構築する。本研究では，2章で説明したエージェントの4つの行動パターンの内，「誘引モード」と「追跡モード」におけるエージェントの行動ルールをQ学習により獲得させることを試みる。具体的には，砂糖を持っていないエージェントがフェロモンを感知した場合の行動を以下のように変更する。

誘引モード，追跡モード；

フェロモンを感知したエージェント i は，時刻 t の環境の状態 $s(t) \in S$ において，Q値に基づいて政策 $\pi^i(s(t), a)$ を決定し，行動を確率的に選択する。

$$\pi^i(s(t), a) = \frac{\exp\left(\frac{Q^i(s(t), a)}{T^i}\right)}{\sum_{b \in A} \exp\left(\frac{Q^i(s(t), b)}{T^i}\right)} \qquad (8)$$

ただし，T^i はエージェント i の温度パラメータである。ここで，もし政策 $\pi^i(s(t), a)$ の値が他のエージェントと同じ値であっても，エージェントの行動は確率的に選択されるため，それらのエージェントが同じ行動を選択するとは限らない。

選択した行動 $a(t)$ を実行すると，環境は状態 $s(t)$ から $s(t+1)$ へ遷移し，エージェントは環境から報酬 $r^i(t+1)$ を受け取る。エージェント i は，状態 $s(t)$，行動 $a(t)$ に対するQ値を以下の式に従って更新する。

$$Q^i(s(t), a(t)) \leftarrow (1-\beta) Q^i(s(t), a(t))$$

$$+\beta^i(r^i(t+1)+\gamma^i \max_{a'\in A} Q^i(s(t+1), a'^i)) \quad (9)$$

ただし，$\beta^i(0<\beta^i\leqq 1)$，$\gamma^i(0<\gamma^i\leqq 1)$ は，それぞれエージェント i の学習率，割引率を表すパラメータである。

5. シミュレーション結果

ここでは，40個のエージェントが50×50の格子状のトーラス平面上に存在する Ant System についてシミュレーションを行った結果について述べる。巣は2次元空間の中心に存在し，ランダムに選ばれた3×3の餌場には1マスに8単位ずつ合計72単位の砂糖が存在する。そして，$\gamma_{dif}=0.1$，$M_p=1$，$\gamma_{eva}=0.1$，$\beta^i=0.3$，$\gamma^i=0.8$ $(i=1, 2, \cdots, 40)$ と設定し，$t_{\max}=2000$ とする100エピソードについてシミュレーションを行った。

まず，上述の設定における初期画面と，50エピソード終了後の画面を図2に示す。ここで，中心に存在する■ (5×5) の領域はエージェントの巣，巣の左下に存在する■ (3×3) の領域はランダムに選択された餌場，右図の餌場から巣にかけて広がっている道筋はフェロモン，黒い点はエージェントを表している。

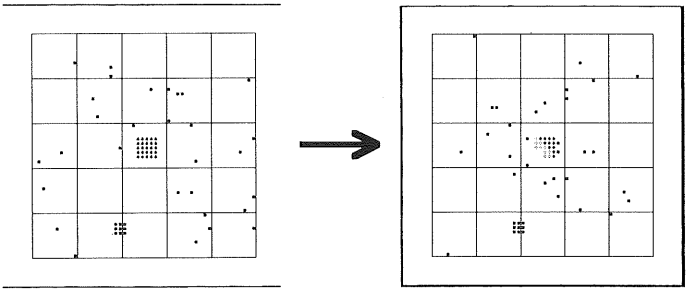

図2　実行画面

図2の右図から，学習を繰り返すことにより，エージェントはフェロモン濃度の高い場所へ移動していることがわかる。これは，従来の Ant System とほぼ同じ結果になっている。

エージェントが試行錯誤的な行動を繰り返し，移動ルールを徐々に獲得する

過程を調べるために，各エピソードにおいて全エージェントが巣に持ち帰った砂糖の合計量（餌所有量）の遷移を図3に示す。ここで，実線は2章で説明した Ant System，太線はQ学習に基づくエージェントによって構成される Ant System の結果を示している。

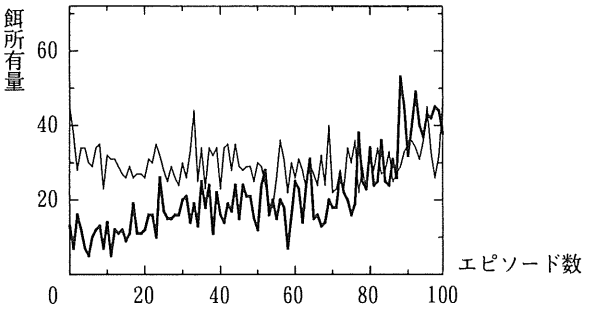

図3　砂糖の合計量の推移

図3から，エージェントがQ学習に基づいて行動する Ant System ではエピソード数が増えるにしたがって，全エージェントが巣に持ち帰った砂糖の合計量が増加していることが分かる。これは，エージェントが初期段階には試行錯誤的に行動していたが，エピソード数が増えるにしたがって環境に対応した行動ルールを確立していることを示している。その結果，最終エピソードでは従来のAnt Systemと同程度の砂糖を収集していることが分かる。

6. おわりに

本論文では，Q学習により誘引モードおよび追跡モードの移動ルールを獲得するエージェントからなる Ant System を構築した。シミュレーション結果から，エピソード数の増加にともない巣に持ち帰った砂糖の合計量が増加している，つまり，エージェントは試行錯誤的な学習を通して，従来法と同様に，フェロモンの量がより多い場所に移動するというルールを進化的に獲得することを示した。今後の課題は，さらに複雑なモデルへ適用し，エージェントの行動を分析することである。

参 考 文 献

[1] 寺野隆雄, 倉橋節也. (2000). エージェントシミュレーションと人工社会・人工経済. 人工知能学会論文誌, Vol. 15, No. 6, pp. 966-973.

[2] 車谷浩一. (2000). 蟻コロニーにおける協調採餌行動のマクロモデルの生成(1)－単純モデルにおけるシミュレーションとモデル生成－. 人工知能学会論文誌, Vol. 15, No. 5, pp. 829-836.

[3] M. Dorigo, G. Di Caro and L. M. Gambardella. (1999). Ant Algorithms for Discrete Optimization. Artificial Life, Vol. 5, No. 3, pp. 137-172.

[4] R. S. Sutton and A. G. Barto. (1998). Reinforcement Learning. An Introduction, A Bradford Book, The MIT Press.