

マルチエージェント系における 強化結果に対する解釈について

白壁啓吾*

乾 岳史**

櫻井 成一郎*

*東京工業大学大学院情報理工学研究科

**株式会社コーエー

自律エージェントを複数協調させることによって問題解決を図るマルチエージェント系の研究が多数行われているが、それらは問題解決の効率向上を主目標としたものであり、組織形成の仕組みについては明らかになっていない。本論文では、特に強化学習によってマルチエージェント系が強調的行動が創発される場合の動的組織形成の仕組みを明らかにすることを目的として、強化学習結果を分析する。本研究で対象とする強化学習結果の分析の道具として、帰納論理プログラミングを用いる。帰納論理プログラミングを分析の道具として用いることで、動的に形成される組織を網羅的に調べることができることが確かめられた。

Explanations of the multi-agents' reinforcement learning results

SHIRAKABE Keigo*

INUI Takeshi**

SAKURAI Seiichiro*

* Department of Computer Science, Tokyo Institute of Technology

** Koei Co. Ltd.

This paper focuses on dynamic formation of collaborative behavior in a multi-agent pursuit game. The dynamic formation of collaborative behavior is achieved by the re-inforcement learning through the interaction with the environment. It is shown that the weighted rules function as triggers for the emergence of commanders which direct the other agents. In such a dynamic formation of a small organization, it is important to share the meaning of the messages. In order to interpret the meaning of the messages, an inductive logic programming(ILP) system is used. By using the ILP system, small dynamic organizations are investigated.

1 はじめに

既存の組織理論では環境と組織形態との適応関係について多くの研究がなされているが、その多くは実証研究であり、組織形態を形成するプロセスについては解明が進んでいない。これに対して、人工知能の分野では、環境との相互作用を考慮したマルチエージェント系の研究成果が近年増加している[3, 4, 5, 6, 7]。個々の適応エージェント(以下エージェント)としては単純な能力しか持たないエージェ

ント群でも、環境中で行動させ、相互に影響を与えながら複雑な組織的行動が創発されるのである。しかし、目標達成という観点から協調行動の創発を研究したものがほとんどで、組織の形成について研究したものはほとんどない。

文献[1]では、強化学習の環境適応性を利用して、自律エージェント間で強調行動をとる場合があることを示している。この研究では、エージェント間で意味共有されていない状態から環境に適応することで意味が共有されて、協調行動がとられる。更に、

回線構造とメッセージ種類数について実験を行い、目標達成率との関連を調べているが、組織形成のプロセスについては明らかにされていない。

本論文では、組織形成においてはメッセージの意味共有が重要な役割を果たすと考え、メッセージの意味付けに対して、帰納論理プログラミング [9] を用いて解釈を付与することで、組織形成のプロセスと意味共有との関係を明らかにする。

2 自律エージェントと組織形成

2.1 自律エージェント

自律エージェントとは、外界から得た情報に基づき自らの判断で行動する個体である。エージェント i が認識可能な状態数を S_i とし、可能な行動の種類を A_i とおけば、エージェント i は $S_i \in A_i$ 個のルールを有するルールベースシステムとしてモデル化することができる。仮にエージェント i の状態 $S_{i,j}$ において、最適と呼べる行動 $A_{i,j}$ が決定できるのであれば、エージェント i は高々 S_i 個のルールを与えておくだけで決定的な振る舞いをするようになる。しかしながら、各状態 S_i 毎にどの行動が最適行動なのかを定めることは容易ではない。そこで具体的に最適行動を定めることなく、環境との相互作用に最適行動の決定を委ねてしまうのが強化学習 [3, 4, 7, 8] である。強化学習によれば、環境適応的な行動の獲得が期待できるわけであるが、環境適応的な行動とは組織的行動なのか、あるいは環境適応的な行動を取る集団を組織と呼べるのかというのが本論文の主たる課題である。

2.2 組織と組織形成

組織形成について議論するためには、組織とは何かということをまず定義しておく必要がある。本論文では、組織とは、所与の目標を共有し、各構成員が自らの役割を認識し、役割を果たすべく行動する集団であると定義する。問題を単純化するために、ある集団の各個体に対して予め目標を与えておき、目標の共有は実現されているものとする。この仮定の下では、ある集団の特定の部分集団が組織となるためには、その部分集団を構成する各個体はまず自らが特定の部分集団の構成員であることを認識している必要がある。次には、特定の部分集団の中で

果たすべき役割を認識し、更に役割を達成するための行動を実行できなければならない。

各個体が自らの所属する組織を認識できるためには、少なくとも組織の構成員間で何らかの情報交換を行う必要があるが、情報交換には構成員間で言語を共有している必要がある。言語が共有されていなければ、たとえ情報交換を行ったとしても行動や意思決定に有用な情報を共有することにはならないからである。構成員間で語彙を共有する事自体も現実には困難であるが、本論文では簡単のために語彙については既に共有されているものとして考察を進めることにする。語彙数を固定しておけば、言語の共有とは、各エージェントが発する語彙の意味を共有することである。自律エージェントは状態に即して行動を決めるのであるから、特定の語彙の発話者にとっては受信者の次の行動を予測できるならば、発話によって受信者の行動を変更したとみなすことができる。逆に受信者側から考えれば、受信した語彙によって自らの行動を決めているのであるから、発話者の意図を理解した上で行動を行ったと解釈することもできるであろう。たとえば、「右へ行け」という語彙を受信した個体が実際に右へ行動するのであれば、観察者からはあたかもある個体が指令を発し、その指令を受けた個体が指令に従って振舞うという動作を観察することになる。

以上の考察から、個体間での情報の流れと特定の情報に対する行動が確定的に決められるのであれば、関係する個体間で組織が形成されたとみなすことにする。但し、ここで注意しなければならないのは、特定の情報に対して特定の行動を採ることは必ずしも知的振る舞いを意味するものではないということである。自律エージェントの持つ知識を各状態に高々 1 通りの行動選択しか許さないようにすれば、明らかに特定の状態に対しての行動は確定的に決定されるが、このような自律エージェントはランダムに行動するよりも目標達成率が悪くなる可能性さえある。したがって、知的エージェント群として振舞うためには、共有される意味の有効性に関する基準を設ける必要がある。有効性の基準を与えることは現実には困難であるので、各語彙に対してどのような機能が割り当てられたのかによって有効性を判定することにする。語彙にどんな機能が割り当てられたのかについては、帰納論理プログラミングシステム [10] を適用することで、語彙に対する解釈を分析する。

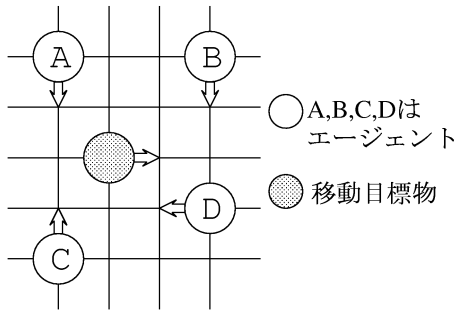


図 1: 2次元追跡問題 [2]

3 学習環境

3.1 2次元追跡問題

本論文では、2次元追跡問題を解決する自律エージェント群を考察の対象とする。2次元追跡問題とは、図1に示すような、無限平面上の自律エージェントが移動目標物を上下左右から取り囲み捕獲するゲームである。エージェント群と移動目標物は2次元格子を交互に移動する。自律エージェント群が移動した後に目標物が移動を完了することを1サイクルとする。自律エージェント群が目標物の上下左右に位置したとき、自律エージェント群は目標物を捕獲する。

本論文の目的は、目標物の捕獲をするための最適なルールを学習することではなく、組織がどのように形成されるのかを明らかにすることであるので、エージェント同士の相互作用を実現するために、エージェント間に通信路を与えておく。文献[1]と同様に、各エージェントは意思決定の場面では、ルールの重みに従ってルールの選択を行うものとし、各サイクル毎に次のメッセージ交換を行う。

1. センサー情報に基づき、参照したルールの重みを付加してメッセージを送信する。
2. 各エージェントは受信可能なメッセージの中で重み最大のメッセージを選び、センサー情報から新たに送信メッセージを決定する。
3. 新しいメッセージを受信しない場合には、直前の送信メッセージを再送する。
4. どのエージェントも新しいメッセージを受信しないか、一定回数のメッセージ交換を行ったら、メッセージ交換を終了する。さもなければ、2に戻る。

メッセージ交換が一定回数(実験では20回)以内で終了する場合には、各エージェントは最大重みを持つルールを選択することになる。各ルールの重みは、強化学習[7]を用いて環境からの報酬を各ルールに分配することで調整する。メッセージに対する重みは行動決定に用いられたルールの重みと参照したメッセージの重みから算出されるので、結果としてメッセージの共有される意味も学習されることになる。

4 学習される行動

4.1 環境からの報酬

環境からの報酬は[1]と同様に、1ゲーム終了時に目標物を捕獲していれば、表1の y を大域的報酬として与える。目標物を捕獲できなかったが、エージェントが目標物に隣接していて、目標位置の競合がある場合には、表1に示す負の報酬 w を与える。エージェントが隣接していない場合には、エージェントと目標物との距離 d を基に、表1の z ($0 < z < y$)の値を与える。大域報酬とは別に、エージェントの各移動毎に、目標物との位置が縮まっていなければ、負の報酬 $-x$ を与える。

これらの報酬の与え方は後述するように、メッセージの意味共有と密接に関連することになる。

表 1: エージェントに与える報酬の設定 [2]

$x = 0:05$
$y = 20:0$
$z = y \cdot E \left(\frac{2}{1 + \exp((d-1)y)} \right)$
$w = 0:05$

4.2 動的組織形成

2次元追跡問題を解くのであれば、各エージェントの移動目標を最初に定めて、その移動目標に向けて移動を繰り返せば、目標移動物の捕獲という目標は容易に達成できる。しかしながら、[2]では実験を繰り返してもそのような傾向は観察できなかった。[2]では、一つのエージェントがメッセージを送信して、受信する部分に着目して、エージェント間で意味共有がなされたと考えられる場合を見出した。すなわ

ち、あるエージェントが他のエージェントに対してメッセージを送信するという事は、自らが認識した役割を果たすために、他のエージェントに対してメッセージを送信していると考えることができる。これは小さな組織形成が動的に繰り返されていく過程であると考えられる。メッセージを送信しても聞いてくれるエージェントがいなければ、メッセージの送信自体がその集団にとって意味のある行動ではなくなるといえる。逆に受信するエージェントが存在して、受信したエージェントの意思決定に影響を与えることができるのであれば、エージェントがメッセージを発信したことが何らかの意味を持つと考えることができる。より踏み込んで考えれば、発信者と受信者の当事者だけに限れば、その意味はより大きくなるのではないかと考えられる。すなわち、メッセージ送受信の当事者が小さな組織を形成していると考えることができるのではないだろうか。そこで、当事者間でメッセージの意味すなわち発信者の意図を共有でき、両当事者がその意図に従って意思決定を行っているという解釈できる場合に、動的組織形成がなされたと捉える事にする。

このように小さな組織が動的に形成されると考えた場合には、如何にして発信者の意図を推し量るかが問題となる。本論文では、観察者が発信者の意図を推し量るための道具として、帰納論理プログラミングを用いることを考える。

5 帰納論理プログラミングによる意味解釈

5.1 帰納論理プログラミングシステム Aleph

帰納論理プログラミング (ILP) は、帰納的機械学習と論理プログラミングを統合した枠組である。ILP システムは、学習対象となる概念の訓練例が与えられ、訓練例が証明可能となる論理プログラムを導出するシステムである。本論文では、メッセージの共有された意味の解釈を与えるために ILP システムの一つである Aleph[10] を用いる。Aleph は Muggleton の Progol[9] を拡張したシステムで、正例だけではなく、負例も利用して学習仮説を生成できる。基本的には、(1) 例題の選択、(2) 最も特殊な節の生成、(3) 汎化、(4) 冗長性の除去というアルゴ

リズムで学習仮説が構成される。本論文における意味解釈には、データを加工した後、Aleph を用いて論理プログラムを生成することで、発信メッセージの意味付与を観察者の立場で試みる。

5.2 強化学習結果に対する Aleph の適用

マルチエージェント系に対する強化学習とは、環境を訓練例あるいは教師とした、選択的学習であって、初期ルール集合の中から環境に適合するルールを選び出すという学習である。但し、単にルールを選択するだけでなく、環境への適合度に応じた重みが各ルールに対して付与される。重みはメッセージ選択やルール選択においては重要な役割をなすが、シミュレーションのトレース結果すなわちルールの利用履歴を利用することで、各ルールの重みについては単純化のために無視することにする。ルールは四つ組 $(s; i; b; o)$ によって表されるものとし、センサー値 s と入力メッセージ i が決まれば、各エージェントは最大重みを持つルールを選択する。

本論文の目的は、環境に適合する組織がどのように動的に形成されるのかというプロセスを明らかにすることであるので、重み付きルールを用いて網羅的に環境を生成して分析することは行わずに、擬似乱数を用いて環境を動的に生成し、その中で形成されるプロセスを分析する。

5.3 Aleph の適用結果

各エージェントが学習した結果を利用して、環境を網羅的に生成した上で、分析するのは計算時間を擁するので、本論文では擬似乱数を用いて環境を動的に生成し、動的環境での組織形成について分析を行う。具体的には、強化学習が完了したマルチエージェント系に対して、追加試行シミュレーションを行うことで、メッセージ交換及びルール選択の履歴を収集する。収集した履歴から ILP に必要なデータを加工して、ILP システムに与える。

ILP のための訓練例としては、エージェント i からエージェント j に対してメッセージ m というメッセージが送られたという事実 $\text{send}(i; m; j)$ を正例とする。なぜそのメッセージが送られたのかを説明する節集合を得るために、背景知識として、シミュレーション時に用いられたルールを与えておく。各ルールは、 $\text{rule}(a; s; i; b; o; t)$ という事実として与え

る。aはエージェント名を、sはセンサー値を、iは入力メッセージを、bは行動目標位置、oは出力メッセージを、tは時刻をそれぞれ表す。行動目標位置とは、目標捕獲物の上下左右のいずれかを表す。

ILPでメッセージの意味解釈を行うためには、解釈となり得る候補仮説を予め与えておく必要がある。本論文では、候補仮説として、行動目標位置の競合解消規則を背景知識として与えて実験した。この競合解消規則は、以下の事実集合により

```
resolv(0; 1):resolv(0; 2):$$$
```

で与えた。更に次の時刻を表す述語 $\text{inc}(X; Y)$ の定義も背景知識として与え、実験したところ、次のような節が得られた。

```
send(A; B; C) : i
rule(A; :-, :-, D0; B; T);
rule(C; :-, B; D1; :-, T1);
inc(T; T1);
resolv(D0; D1):
```

この節は、

エージェント A からエージェント C にメッセージ B を送る理由は、エージェント C に自らの行動目標との競合解消をしてもらいたい。

と読むことができる。これはマルチエージェント系が強化学習によって上記知識を学習したと考えることもできるし、エージェント A と C はメッセージ B を互いに「行動目標の競合解消」という意味であると認識しているとみなすこともできる。その結果、エージェント A と C は動的に組織を形成したと考える。

5.4 ILP の適用結果に関する考察

前節で述べたような意味共有の解釈を行うために ILP を適用する際の問題点は、解釈候補を予め与えておかなければならないということである。ILP システムで必要な概念を構成することも可能ではあるが、有用な補助概念を構成することは極めて困難である。本論文で与えた解釈候補は、強化学習の際の報酬の与え方から推定したものである。4.1 節で述べ

たように、一定回数の移動の後、目標物を捕獲できない場合には、移動目標の競合に対して負の報酬を与えているので、この負の報酬の効果が競合解消規則の学習に寄与しているのではないかと考えられる。

意味共有の解釈を行う目的は、組織形成プロセスの解明であるが、直接組織形成のプロセスを知識として ILP で得る事はできなかった。例えば、組織形成の分析の立場であれば、以下のような節を直接導出できることが望ましい。

```
send(A; B; C) : i
役割認識 (A; D0; B);
役割 (C; D0):
```

しかしながら、「役割認識」や「役割」などの概念を背景知識として与えるのが困難であったので、導出することができなかった。マルチエージェント系のルールとしては、ルールの重みが相対的に大きい場合に、「役割認識」の機能を果たしていると予想されるが、重みを無視したために、このような直接組織形成を説明する知識を得る事はできなかった。

ILP を適用する利点は網羅的に仮説の妥当性を検証する事であり、膨大な数のルールの中からどのような意味共有が行われているのかを調べるには有用であることが確かめられた。しかしながら、エージェント毎にルールセットが異なり、また各時刻毎にルールを識別する必要があるので、大量の試行データから直接 ILP を行うのは困難であったから、データについては前処理を行わざるを得なかった。これらの前処理を自動化し、他の意味共有が行われているのかどうかの分析を行うのは今後の課題である。

6 おわりに

本論文では、自律エージェント群が環境からの報酬に基づく強化学習を行っていくことでどのように組織が形成されるのかについて考察し、考察結果を検証するために ILP を用いて強化学習結果に対する解釈を与えた。その結果、この単純な実験においても、エージェント間でのメッセージの意味共有が実現されていることを検証することができた。

組織形成プロセス解明のための道具として、ILP を利用するためには、解釈候補となる知識を背景知

識として与えておかなければならない。解釈候補となる知識をどのように与えるかということは今後の課題である。また、生データを ILP で利用するためには、変数の型情報や知識の依存関係も予め与えておかなければならないが、現在は手作業で与えている。これらの前処理を自動化することで、より詳細に組織形成のプロセスを解明することは今後の課題である。

参考文献

- [1] 松浦・嘉数：非均質 agent 系における組織的行動の形成: 2次元追跡問題における考察, 情報処理学会論文誌, Vol. 38, No. 6, pp.1083-1093, 1997.
- [2] 乾・櫻井：マルチエージェントによる組織形成に関する基礎的研究, 情報処理学会研究報告 (2001-GI-5), pp.23-30, 2001.
- [3] 畝見：強化学習「最近の機械学習」, 人工知能学会誌, Vol.9, No.6, pp.830-835, 1994.
- [4] 畝見：強化学習エージェントの集団行動, MACC'93, pp.137-150, 1993.
- [5] 大沢：協調プランニングの動的組織再編とメタレベル整合戦略-追跡ゲームにおける考察-, コンピュータソフトウェア, Vol.12, No.1, pp.511-519, 1995.
- [6] 大沢：問題空間の変化に適応する協調的組織スキーマ-追跡ゲームによる考察-, MACC'92, pp.105-120, 1992.
- [7] 荒井・宮崎・小林：マルチエージェント強化学習の方法論-Q-Learning と Profit Sharing による接近, 人工知能学会誌, pp.609-617, 1998.
- [8] 山村・宮崎・小林：エージェントの学習 人工知能学会誌, Vol. 10, No.5, pp.683-689, 1995.
- [9] S. Muggleton: Inverse Entailment and Progol, New Gen. Comput., Vol. 13, pp. 245-286, 1995.
- [10] A. Srinivasan: Aleph, http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph_toc.pl