人　工　知　能 93−10
（1994．3．8）

学習能力を持つ自律エージェントによる協調的行動

坂本 忠昭

三菱電機 (株) 中央研究所

本稿では、分散 AI の標準的小問題の 1 つである追跡問題を例題として、自律エージェントに協調的行動を自発的に生成させる試みについて述べる。エージェントは 2 次元平面上を動き回り、捕食型のエージェントが被食型のエージェントを追いかけ捕まえるものと仮定する。捕食型のエージェントは、局所的な視覚情報とマッチする行動規則に従って次の行動を決定し、行動結果の評価を通して行動規則を生成し洗練する学習能力を備えている。我々の目的は、このような学習機能を持つ捕食型のエージェントの集団が、囲い込みのような協調的な行動を生成するためのメカニズムを探ることである。

Cooperative Behaviors of Autonomous Agents with Learning Ability

Tadaaki Sakamoto

Central Research Laboratory, Mitsubishi Electric Corp.
8–1–1, Tsukaguchi-Honmachi, Amagasaki, Hyogo, 661, Japan
sakamoto@sys.crl.melco.co.jp

In this paper, we present our trial to make autonomous agents emerge cooperative behaviors using one of canonical problems in distributed AI, “pursuit game”. All agents move around on 2-dimensinal plane, and predator-type agents chase prey-type agent to capture it. Predator-type agents decide next action according to the behavior rule which matches its local visual information, and have a learning ability to create and refine rules through an evaluation of an action and/or a sequence of actions. Our aim is that coorperative pursuant behavior like a surroundings will be emerged by a group of predator-type agents.

1 はじめに

人工生命における最も重要な概念の1つは創発 (emergence) である。これは、系の複雑な振舞いは、その系を構成する数多くの単純な要素が局所的なインタラクションを行うことによって自発的に生成される、という考え方である。この概念を採り入れることにより、一見複雑に見える生命現象も実際にはもっと低レベルの構成要素の集団的振舞いという形式にモデル化できる、あるいは、複雑な問題も単純な問題解決器群によって解決することができる、といった期待が持たれる。そして、これらの目的のために、エージェントの集団行動に関する研究が盛んに行われている。

ただし、多くの場合には、各エージェントの行動規則は人間が予め設計し与えており、研究の対象は、エージェントにどのような行動規則を持たせるとどのような集団行動パターンが生成するかということである。エージェント同士の協調的行動に関しても、他のエージェントと協調するための行動規則は予め与えられていることが多く、それをエージェント自らが獲得していくという研究はまだあまりなされていない。

個々のエージェントの行動に関しては、環境への適応というテーマで研究が行われており、行動規則を予め与えなくても、エージェントが環境とのインタラクションを通して行動規則を次第に獲得していくという学習方法が提案されている。このような学習方法としては、分類システム＋遺伝的アルゴリズム ([5] など)、ニューラルネットワーク＋遺伝的アルゴリズム ([1] など)、Lisp プログラム＋遺伝的アルゴリズム (Genetic programming) [2]、実例に基づく強化学習 [6] などがあげられる。本研究の目的は、このような適応学習の枠組を用いて、エージェントの集団に協調的行動を学習させることである。

本稿の構成は以下のようになっている。まず、関連研究としてマルチエージェントの強化学習の研究を紹介し、本研究との比較を行う。次に、対象問題である追跡問題について説明し、続いて、我々が学習方法として用いた実例に基づく強化学習について述べる。そして、その学習方法を用いたエージェントの設計を行いながら幾つかの問題点を明らかにし、その検討を行う。

2 関連研究

本研究の関連研究として、Tang による複数のエージェントにおける強化学習の研究 [4] がある。そこでは、10×10 のグリッド平面上で、Q-learning アルゴリズム [7] を用いた強化学習機能を持つハンターエージェントが、ランダムに動き回る獲物エージェントを追いかけるという設定の追跡問題を例題とし、以下の3つの場合について協調的行動による効果を調べている。

(1) 感覚の共有

獲物、ハンターに加えて、偵察者を仮定する。偵察者はランダムに移動しながら各ステップ毎に自分の行動と視覚情報をハンターに送る。ハンターは自分の視覚情報に加えて、偵察者からの視覚情報を利用して獲物を追跡する。偵察者の存在によって獲物を捕獲するまでの時間は短縮され、さらに、2つのハンターがお互いに相手の偵察者として働くことによって、この効率は向上する。

(2) 意思決定やエピソードの共有

あるハンターが意思決定機構を持ち、全てのハンターがその機構を用いて意思決定を行う場合と、一定間隔でハンターがお互いの意思決定規則を交換し、自分と相手の意思決定規則を融合していく場合を設定する。両者とも各ハンターが独立に意思決定を行う場合よりも学習効率が向上するが、両者の間に大きな差はない。さらに、あるハンターが獲物を捕らえたときに、捕らえるまでのエピソード

（視覚情報、行動、報酬の3つ組のシーケンス）を他のハンターに教えるという場合を設定する。この場合は、エキスパートからエピソードを教えられたときに学習効率は大きく向上する。

(3) 共同作業

(1),(2) と異なり、2つのハンターが1つの獲物を挟まなければ捕獲できないという設定にし、ハンターが相手を視覚情報に入れる場合と入れない場合、そしてお互いに相手の偵察者になる場合の3つを設定する。ハンターが相手を見れない場合はほとんど学習効果がないのに対し、他の2つの場合には学習効果が現われている。また、獲物が2つの場合にはお互いに相手の偵察者になる方が学習効率が高いが、獲物が1つの場合には両者にほとんど差は見られない。

この研究では、本来の強化学習の枠組に協調のためのメカニズムを追加する方法がとられている。しかし、意思決定機構を特定のエージェントだけに持たせるのは、自律エージェントの定義に反すると考えられるし、意思決定規則を交換したりエピソードを教え合うのは、学習効率を高める可能性はあるが、その反面、規則の画一化を招くことも考えられる。その中で、Tang が passively-observing と呼んでいる、ハンターが仲間を視界情報の中にとり入れることのみによって協調的な行動を学習する方法があるが、これは強化学習の枠組をそのまま用いた非常に素直な方法であるため、我々もこの方法を用いることにする。

さて、この研究では既に passively-observing 法を用いて協調的な行動を学習しているが、これは例題の質によるものが大きいと考えられる。まず、例題としている追跡問題の世界は 10 × 10 のグリッド平面であり、絶対的に狭い。反面、ハンターや偵察者の視界は深さ2～4であるが、これは平面上の 5×5 ～ 9×9 の領域を意味しており、相対的に広い。さらに、獲物はランダムに動いていることを考え合わせると、未熟なハンターが獲物を捕らえる確率は高く、強化学習がうまく行われたと考えられる。そこで、我々は例題の質を上げ、その中でうまく学習ができるかどうか調べてみることにした。

3 追跡問題

分散人工知能の標準的小問題 [3] の1つとして、Benda らによる追跡問題がある。これは複数のエージェント間における協調のための組織的関係や通信の影響を調べるために提案された問題であるが、何人かの研究者によって、協調を行うための組織化手法の有効性を確認するために用いられている。

オリジナルの追跡問題は以下のようなものである。まず、無限に広がる2次元のグリッド平面を仮定し、その上を移動するエージェントを考える。エージェントはグリッド上を前後左右方向に1回に1マス移動できる。エージェントには逃げる側の赤いエージェントと、追跡する側の青いエージェントがあり、平面上に赤が1つ、青が4つ置かれている。青いエージェントは赤いエージェントを見ることはできるが、仲間の青いエージェントを見ることはできない。ただし、仲間の青いエージェントの位置は、通信機能によって得ることはできる。このような状況の中で、青いエージェントたちが協力して赤いエージェントを取り囲むことがこの問題の目的である。

これに対し、我々はより現実世界に近い問題設定を行うため、幾つかの点で変更を加えた。まず、2次元グリッド平面ではなく実数平面とした。各エージェントは限られた視界を持ち、その中に入ったエージェントを種類に関係なく見ることができるものとした。また、移動方向はエージェントの現在の方向を中心としたある扇型の範囲内とした。そして、逃げる側のエージェントはランダムに移動するのではなく、追いかけるエージェント

を発見した場合にはそれから逃げようとする行動をとることにした。これらの変更によって、Tangの例題よりもかなり難しい問題設定となる。

4 実例に基づく強化学習

実例に基づく強化学習は、自律エージェントの環境への適応学習アルゴリズムとして畝見によって提案された学習方法である[6]。

学習するエージェントとして、2次元実数平面上を移動する虫を仮定している。虫は扇型の視野を持ち、その中心角を10等分した細い扇型の範囲に見える物体の種類と距離に応じた値を要素とする10次元ベクトルを視覚情報として持つ。虫が見る物体は餌とゴミであり、視野の最も近くにある餌は1、ゴミは-1、視野の最も遠くにある餌は9、ゴミは-9で表される。何も見えない場合は0となる。

視覚情報が入力データとして与えられると、意思決定機構が対応する出力データを決定する。出力データは虫の移動方向の変化分を表す行動データであり、$\{-45°, -30°, -15°, 0°, 15°, 30°, 45°\}$のいずれかの値をとる。意思決定機構は、過去に記憶された入力データの中から、現在入力された視覚情報と最も類似度が高くかつ強化入力の大きいものを検索し、そのときの出力データを出力データとする。出力データによる行動の結果、虫が餌にぶつかるとそれを食べることができたとして強化入力$+1$が報酬として与えられ、ゴミにぶつかった場合には-1が罰則として与えられる。何もない場合は0である。虫は入力データ、出力データ、強化入力の組みを時系列で記憶しており、ある時点で与えられた強化入力は、減衰しながら過去へ遡って伝播するようになっている。これは、ある時点での行動結果は幾つもの行動規則の連鎖によって生じるものであり、報酬も罰則も過去に遡って関係した規則に与える必要があるためである。ただし、過去へ遡るほど現在への影響は弱いという仮定で、伝播の際に減衰をさせている。そして、虫が動き回って経験を積みながら、強化入力の低いデータや、記憶の古いデータを削除していくことにより、次第に頻繁に正の強化入力を得る、すなわち、頻繁に餌を食べるような行動を行うようになるというものである。

この学習方法の利点は、視覚情報と行動データのペアが直接行動規則となり、経験と共に行動規則が増え洗練されていくという点にある。現在見ている状況と良く似た状況が過去にあり、そのときの行動結果が割と評価の高いものであった場合、今回もその行動をとれば良いという考え方は、生物の行動アルゴリズムとして自然であるし、また理解しやすいものである。また、時系列というものを学習アルゴリズムや行動規則記憶で陽に扱っている点も、獲得したい協調的行動パターンが時系列上に現われるものであることを考えた場合、この方法が有効と考えられる理由の1つである。

5 エージェントの設計

2次元平面上を移動するエージェントとして、逃げる側のエージェント（被食者と呼ぶ）と追いかける側のエージェント（捕食者と呼ぶ）の2種類を用意する。以下、各エージェントについてその基本的な仕様を述べる。

5.1 被食者

被食者は学習機能を持たないエージェントであり、決まったアルゴリズムに従って機械的に行動する。しかし、ランダムあるいは一定パターンで動くものではなく、捕食者の動きから次の動作を決定する。

まず、被食者は扇型の視野を持つ。その半径と中心角は初期設定によって自由に変えられ、360°の視界を持つことも可能とする。被食者は、視界の中にいる捕食者の位置と移動

方向からその捕食者の次の時刻の位置を予測し、その予測位置から遠ざかる方向に自分の移動方向を変更する。複数の捕食者が視界の中にいる場合には、個々の捕食者に対して同様の計算をし、得られた方向を合成する。なお、自分の移動方向を変更する場合、変更できる最大角を初期設定によって制限することもできる。これによって、急に向きを反転したりできなくすることも可能である。

さらに、捕食者の協調的集団行動の生成を実験するために、複数の捕食者が、例えば対象方向から向かってくるなどある集団行動パターンをとる場合に移動速度を減少させ捕まりやすくなるという設定が必要である。

5.2 捕食者

捕食者は行動規則に基づいて行動し、その行動規則を学習する機能を備えている。まず、図 1に捕食者の行動規則の表現方法を示す。行動規則の条件部は視覚情報である。捕食者の視界は図 1に示すようなメッシュ状に区切られた扇型であり、各セルに位置を示す番号が割り当てられている。図では、5の位置に被食者 (種類1)、6の位置に捕食者 (種類2) が見えている場合を示しているが、この場合には視覚情報として 000001200 が得られる。この方法は [6] における視覚情報の表現と異なるが、メッシュの同じセルに入らない限り異なるエージェントを常に認識できる、あるいは、認識可能な物体の種類が増加した場合には対応する番号を増やすだけでそのまま対処可能であるといった利点を持つ。なお、図では3×3のメッシュを示しているが、視界を幾つのメッシュに区切るかは初期設定によって変更可能である。また、エージェントが選択可能な移動方向は図に示されるように番号付けされており、この番号が行動規則の行動部として用いられる。この移動方向の角度や数も初期設定によって自由に設定可能である。

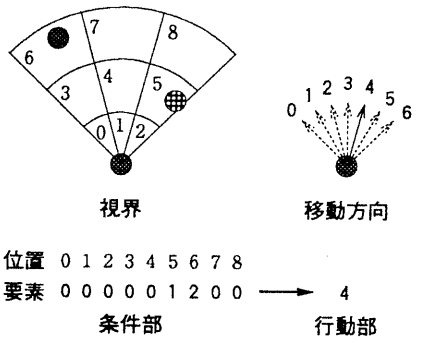

図 1: 行動規則の表現

次に捕食者の内部構成を図 2に示す。行動決定部は視覚情報から行動データを決定する部分である。視覚情報とマッチする条件部を持つ行動規則がルールベースにあれば、その行動部が行動データとなる。条件部は、視覚情報が1である位置が1、2である位置が2となっていればその視覚情報とマッチしたと判定される。以下に簡単な例を示す。

```
視覚情報： ０００００１２００
----------------------------
条件部１： ０００００１２００  match
条件部２： ０００００００２００  unmatch
条件部３： ０００００００２１０  unmatch
条件部４： ００００１１２＃０  match
```

条件部1は視覚情報と全く同じためマッチする。条件部2や3の例では、視覚情報が1である位置が0になっているためマッチしない。条件部4はマッチングの条件を満たしているためマッチする。ここで、＃は don't care を表す。条件部4がこのような形を持つのは、これが後で述べる行動規則の一般化の結果生成される条件部を示しているためである。なお、この例では、与えられた視覚情報に対して、条件部1と条件部4がマッチするが、そのときには視覚情報との差分の少ない条件部1が選択される。

マッチする行動規則がなければ、ルール生成部で新たな行動規則が生成される。新しい

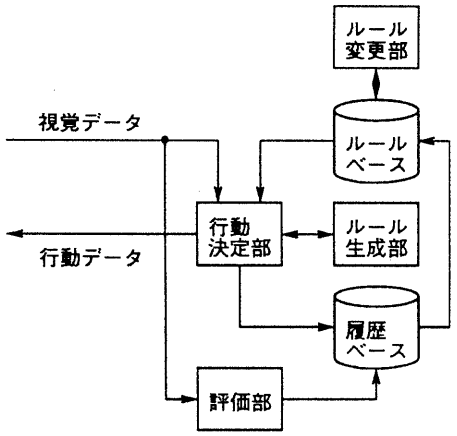

図 2: 捕食者の内部構成

規則は、視覚情報を条件部にし、ランダムに生成された行動データを行動部にする。新たに生成されたものも含めて、用いられた行動規則は履歴ベースに時刻順に蓄積される。履歴ベースは有限長であり、それを超えた場合には古い履歴から順に削除されるものとする。そして、被食者を捕まえたり逃がしたりといったあるタイミングで、履歴ベース内の行動規則はルールベースに送られる。このとき、既にルールベース内に存在する行動規則については、評価値の更新のみが行われ、新しく生成された行動規則はルールベースに新たに蓄積される。ただし、既にルールベースが一杯であり、自分よりも評価値の小さい規則が存在しない場合には、その規則は破棄されることになる。

評価部は視覚情報に基づき強化入力を生成する部分である。得られた強化入力は履歴ベースに送られ、蓄積されている行動規則の評価値として用いられる。評価としては、短期評価と長期評価の2種類を考えている。短期評価は直前に適用された行動規則に対してのみ与えられ、長期評価は過去に遡って複数の行動規則に対して与えられる評価である。現在のところ、短期評価としては、被食者に接近すれば正、遠ざかれば負の値を与えるもの、長期評価としては、被食者を捕まえれば正を与えるものや一度視界に入れておきながら逃がした場合には負を与えるもの等を検討中である。本来の強化学習の枠組では、被食者を捕まえたときに強化入力 +1 を与えるだけであるが、我々の例題では被食者の逃げ方がうまいため、これではいつまでたっても +1 を獲得できないと考えられる。そのため、獲物に接近するといったような、短期的な目標を達成することによって得られる評価値も必要になってくると思われる。

ルール変更部は行動規則の一般化を行うための部分である。[6] では、実例に基づく強化学習は、訓練例を加工せずにそのまま記憶し、類似度計算によって検索することを特徴とする学習方法であるとされている。しかし、次節でも述べるように、視覚情報のパターン数が増加した場合の行動規則の一般化は、ルールベースの大きさをある程度に抑えるためには必要であろう。ここでの一般化は、同じ行動部を持つ2つの規則を任意に選択し、その条件部の論理和を求め、それを新しい条件部にした規則を新たに生成することによって行われる。ある位置において一方の要素が1他方が2の場合には、# (don't care) が新しい要素となる。元の2つの行動規則はそのまま残され、新たに生成された規則がルールベースに登録される。

我々の狙いは以下のようなものである。被食者は捕食者が取り囲むようにやってきたときに速度が落ち、捕らわれやすくなるようにする。すると、捕食者が被食者を捕らえ、強化入力を得たとき、その捕食者の視界には仲間の捕食者が入っているだろう。逆に、仲間の捕食者が被食者を捕まえるのを見た場合にも強化入力を得るようにしておけば、捕食者達はお互いに今までの行動規則集合、すなわち一連の行動パターンが捕獲に関して効果の高いものとして記憶していくだろう。その結

果、捕食者達の中で被食者を捕獲するためのフォーメーションのようなものが形成されるのではないかというものである。

6 主な課題

前節の仕様に従って、現在 SPARC station 2 上に実装中である。実験段階で幾つかの課題や問題点が見つかっているので、ここではそれらについて述べる。

(1) 視覚情報の大きさと精度

被食者・捕食者の両方が動き回るため、視覚情報は時間毎にかなり激しく変化する。加えて、被食者が仲間を視野に入れる必要があるため、視野の半径や角度は大きく、そしてメッシュ数は多くとることが必要と考えられる。さらに、被食者が動かない場合には、視覚が認識するものは物体の種類だけでもよいが、被食者が動く場合にはその移動方向も認識する必要があるのかもしれない。ところが、これらの条件を満たしていくと、視覚情報の状態数が飛躍的に増大することは明らかである。状態数の増大は、記憶容量の増大や学習速度の低下等の問題を伴うことが予想されるため、これを抑える工夫が必要になってくる。

(2) 行動規則の一般化

視覚情報の状態数の増加に対応する方法として、情報の一般化がある。現在のところ、先に述べたように、ルール変更部において行動規則条件部の一般化を行う方針である。ただし、帰納学習においても問題になっているように、与えられた事例をどの程度一般化するかというのは難しい問題である。ただむやみに一般化していったのでは、何にでも適用できる役に立たない規則になってしまう。どのタイミングでどの規則をどの程度一般化するかというのは重要な課題になるであろう。

逆に一般化を行わないとすれば、視覚情報の状態数の増加に対応する他の方法が必要になってくる。

(3) 強化入力の与え方

被食者の逃げ方が上手である場合には、捕食者が適当に動いたのでは捕獲するチャンスはほとんどない。従って、被食者を捕らえたら強化入力 +1 を与えるという単純な方法では、いつまでたっても強化入力が与えられず学習が進まない。そのため、我々は短期評価、長期評価として幾つかの評価を導入し、被食者に接近することによっても何らかの評価値を受けられるようにしている。しかし、このように複数の評価基準を導入した場合には、それらの評価値の整合が問題となる。

また別の方法として、学習状況に応じて目標を変更していくという方法も考えられる。すなわち、まず初期の目標は被食者にある程度接近することとし、その距離まで接近できれば強化入力 +1 を与える。ある程度学習が進んだところで、目標とする距離を小さくする。このようにして、最終的には被食者を捕らえたときに強化入力を与えるようにしていくのである。しかし、この方法も、目標の切替えをいつ行うかという問題や、このサイクル 1 回で学習が収束する保証がないという問題が残される。

7 おわりに

本稿では、実例に基づく強化学習機能を持つ自律エージェント集団に協調的行動を自発的に生成させる試みについて述べた。協調的行動というのは、個々のエージェントが行った行動のシーケンスの組合せであり、その組合せが全体として効果的なものである。我々の例題では、個々の捕食者が被食者に接近するとき、被食者を囲むように周りから接近するような行動パターンが、被食者を捕まりやすくするという点から協調的行動になるであろう。しかし、検討を進め実験を行うにつれ、適当に動き回っていても強化入力が得られる、といったような問題設定でない場合に

は、各エージェントが有効な行動のシーケンスを学習するということはかなり難しいということがわかってきた。先に述べた課題も含めて、これをどのようにして克服していくかということが今後の課題である。

その先の課題としては、捕食者の移動速度や視野半径に個体差を付けたときに、追いかけ専門や待ち伏せ専門といった役割分担が起こらないかを試してみることや、逃げる側のエージェントにも学習機能を持たせ、両者の戦略がどのように変化していくか調べてみることがあげられる。

参考文献

[1] Collins, R.J. & Jefferson, D.R.: Ant-Farm: Towards Simulated Evolution, *Artificial Life II,* pp.579–601, 1992.

[2] Koza, J.R.: Genetic Programming: On The Programming of Computers by Means of Natural Selection, MIT Press, 1992.

[3] 大沢 英一, 沼岡 千里, 石田 亨: 分散人工知能における標準的小問題, コンピュータソフトウェア, Vol.10, No.3, 1993.

[4] Tang, M.: Multi-agent Reinforcement Learning: Independent vs. Cooperative Agents, *Proceedings of the Tenth International Conference on Machine Learning,* pp.330–337, 1993.

[5] 上田 雄悟, 堂田 敏文, 星野 力: ゲーム環境における分類システムと遺伝的アルゴリズムの学習効果, 人工知能学会全国大会（第7回）論文集, pp.173–176, 1993.

[6] 畝見 達夫: 実例に基づく強化学習法, 人工知能学会誌, Vol.7, No.4, 1992.

[7] Watkins, C.J.C.H. & Dayan, P.: Technical Note: Q-Learning, *Machine Learning,* Vol.8, No.3/4, 1992.