第4回AI若手の集い(人工知能学会主催)
MYCOM2003資料
2003年5月26日，27日

学習空間のモジュール間インタラクション

*山岸栄輝，塩瀬隆之，川上浩司，片井修

京都大学大学院情報学研究科
〒606-8501 京都市左京区吉田本町
Phone: (075)753-3592, Fax: (075)753-5042
{yama, shiose, kawakami, katai}@symlab.sys.i.kyoto-u.ac.jp

Abstract: 強化学習の一手法である Q 学習は、エージェントが未知環境において試行錯誤を通じて行為を学習する枠組みとして注目されている。Q 学習では、タスクに関する状態変数とエージェントの行為変数の組み合わせ、つまり＜状態・行為＞の組に対する評価関数を更新することによって、適切な行動を学習する。ここで、状態変数と行為変数の数が多ければ、＜状態・行為＞の組み合わせ（学習空間）が爆発的に大きくなり学習が困難になるという問題を抱えている。そのため、これまで学習空間の構成方法についてさまざまな提案がなされている。たとえば、学習空間の自律生成、モジュール化などである。さらに、モジュール化の中でも、タイルコーディング、階層型モジュール、並列型モジュールなどがある。本発表では、学習空間のモジュール化の研究の一つとしてわれわれの研究を紹介し、モジュール間のインタラクションについて考察する。われわれは、＜状態・行為＞の組に対する評価関数の表現として状態変数のすべてではなく一部を考慮する学習空間モジュールを複数組み合わせて用いることによって、学習空間の低次元化を実現している。

1 はじめに

未知環境や動的環境など複雑な環境において知的な振る舞いを実現する手法の一つとして、試行錯誤を通して自律的に行動を学習する強化学習の枠組み [1] がある程度成功を収めてきた。そして、この強化学習をより動的で複雑な環境に適用しようとすると学習すべきパラメータが通常多くなるため、学習を効率化するため学習空間をモジュール化する研究が数多くある。モジュール化によって学習を役割分担し、それと同時に環境変化に対する頑健性も期待されている。

ここで、われわれが問題にしたいことは、モジュール化の仕方についてである。つまり、複雑な環境に適応するよう人間があらかじめ役割を決めたモジュールを作成し埋め込むことは困難なのではないかということである。もちろん適用する問題の性質にもよるが、人間の設計したモジュールの役割は必ずしも未知環境、動的環境で有効に働かない可能性があり、また、そもそもモジュールの役割が人の用意した役割に落ち着かなければならないという必然性はない。

環境の変動に対して頑健なシステムは、人間の用意していない役割が状況に応じて動的に決まってくるようなものであると考える。環境の状況に応じて適当なモジュールが役割を担えるようなシステムを考えることが重要ではないだろうか。そのようなシステムを考えるために、自然農法とシードバンクについて紹介する。

2　自然システム

2.1　自然農法

福岡正信によって提唱された自然農法は、「無耕耘（むこううん）」「無肥料」「無農薬」「無除草」で特徴づけられる農業のあり方である [2]。そこでは、様々な種類の種子を粘土団子の中に同時に混ぜ入れ、様々な土地（草地、荒地、林、農地）にばら播く。その中から何の種子が発芽するかは自然の摂理に任せられており、多種多様な動植物が混生する農園となる。その結果、一つの野菜に対して害虫となる虫が、他の野菜にとっての益虫となる場合がある。また、たとえ害虫が存在してもその天敵が存在すれば系としてのバランスが保たれ、大きな被害を被ることはないという。つまり、野菜や果物、稲などを生態系の中に組み込み栽培すれば、上に挙げた 4 つの事を実践しても収穫を得ることができる。

2.2　シードバンク

樹木や草は子孫を残すため、種子を風、鳥、動物や人など、様々な方法で種子の落ちる範囲を広げる。これらにより運ばれた種子は、地上に落ち、落葉の下や雨により土中へと入ってゆく。種子はその環境により発芽するもの、発芽せずに死ぬもの、また、土の中で休眠し何年も生き続けるものもある。土の中で休眠する種子は、森林では山火事や倒木によりその空間の陽当たりが良くなった場合、また、都市部では工事により土を堀起こし他へ運びだした場合など、今までと環境が変わる場合に発芽することがある。このように、土の中に休眠している種子を「埋土種子」と呼び、土の中に種子が蓄えられていることを、「シードバンク」と呼ぶ。植物はこのようにして環境の状況に適応したときに初めて発芽し、その子孫を残すことができる。

2.3　自然システムの特徴

以上、この章で言いたいことをまとめると次のようになる。未知環境や動的環境など複雑な環境において、適切に役割分担がなされ頑健な振る舞いを学習するモジュール構成は、

1. 様々な種（モジュール）を準備して仕込んでおくこと
2. 様々な動植物や自然環境（他のモジュール）とのインタラクションで、適材適所的に種（モジュール）が生長すること
3. 他の植物や環境（モジュール）との関係が適せず発芽しない種（モジュール）はシードバンクに保存され（未発達のままのモジュールが存在し）、時が来れば発芽すること

が重要である。また、モジュールの役割については、

1. あるところでは害を及ぼす虫や雑草（モジュール）も、別の虫や植物（モジュール）にとっては重要な働きをすることから、虫や雑草（モジュール）の役割は時と状況によって、また、役割を判断する主体によって異なること
2. 複雑な環境においては、役割をあらかじめ決定することは困難であること

を意識することが重要である。

3　学習空間のモジュール化

前章で見たことは、未知環境や動的環境に適応するため学習空間をモジュール化する場合、あらかじめモジュールの役割を決めず、ただ役割分担が生じるような様々なモジュールを仕込んでおき、どのような役割が生じるかはモジュール間のインタラクションに任せるということであった。そこで、われわれのアプローチを紹介する。

3.1　部分的条件省略ファジィルールを用いた Q 学習

機械学習の一手法である Q 学習は、試行錯誤を通じて、エージェントにとっての＜状態・行為＞の組合せに対する評価値 (Q 値と呼ぶ) を学習する枠組である [3]。ここで、＜状態・行為＞の組からその評価値 (Q 値) への写像を Q 関数と呼び、もっとも単純な場合、表形式を用いて Q 関数が実現される。

われわれは、Q 関数の表現にファジィ推論を導入し連続値入出力を扱うことができるファジィ内挿型 Q 学習の学習空間を複数に分割する “部分的条件省略ファジィルールを用いた Q 学習” を提案し、学習の効率が良くなることを確認している [4]。部分的条件省略ファジィルール (CRFRs: Condition Reduced Fuzzy Rules) における前件部は、エージェントの状態を表す変数の一部とエージェントの行為を表す変数に関する条件で構成されている。つまり、図 1 に示すとおり、対象問題を解くために十分と思われる n 個の状態変数のうち N_r 個 ($N_r < n$) 個の状態変数と、エージェントの行為変数との関係を学習するモジュールを ${}_nC_{N_r}$ 個用意して、N_r 個の変数と行為変数との関係を網羅的に学習するようにモジュールを構成する。本手法の学習アルゴリズムを表 1 に示す。

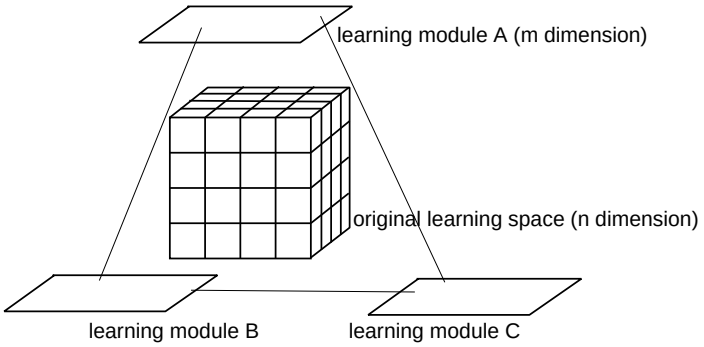

図 1: 学習空間のモジュール化

各モジュールは、1) それぞれの持つファジィルールと＜現在の状態、仮定された行動＞の組との一致度、2) その組に対する評価値の 2 つを見積もる。そして、各モジュールから計算された一致度と評価値からファジィ推論によって、＜現在の状態・仮定された行動＞に対する最終的な評価値が推論される。このようにして、複数個の仮定された行動に対する評価値を見積もり、実際に行為として実行される行動を確率的に選択する。

表 1: 提案手法の学習アルゴリズム

1. すべてのファジィルールにおける＜状態、行為＞の組に対する評価値 C_i を初期化
2. Repeat Forever
   (a) Repeat T times
      i. 行動をランダムに仮定
      ii. ＜現在の状態、仮定された行動＞の組に対する各ルールの一致度 ω_i を求める
      iii. 現在の状態と仮定された行動の評価値を見積もる
   (b) (after second cycle)
      i. 前回の＜状態、行為＞の組に対する評価値 Q 値の更新量 ΔQ を以下の式で求める

$$\Delta Q = \alpha\{r_t + \gamma \max_b Q(x_{t+1}, b) - Q(x_t, a_t)\} \tag{1}$$

      ii. 次の式にしたがって ΔQ を各 CRFR の評価値の更新量 ΔC_i に分配する

$$\Delta C_i = \frac{(\sum_j \omega_j \omega_i \Delta Q)}{(\sum_j \omega_j^2)} \tag{2}$$

   (c) 仮定された行動のうちから一つを選択して実行する。
   (d) 次の状態と強化信号を観測する

3.2 学習空間モジュール

われわれの提案手法の特徴は、学習空間をモジュール構成する際、性質の多少異なるモジュールを多数用意しておくことである。つまり、これらのモジュールは、自然農法のアナロジーで言えば、人間の用意する「種」に該当する。

次に、モジュールの役割は、各モジュールから出力される＜状態、行為＞に対する一致度と評価値の差異によって生み出される。つまり、環境とのインタラクションを通して環境から得られる報酬によって、各モジュールが徐々に環境に適応し、環境の状態に応じて大きな評価値を出力するモジュールが変化することが予想される。

4 実験

4.1 実験環境

今回は、環境として図 2 に示すコースを周回するような船の操舵問題を取り上げ、あるタイムステップで一つのセンサを故障させることをもって、環境の変動とした。船の操舵問題は、時間遅れが大きいため学習が難しいとされている。

各モジュールが認識できる状態変数は、船の位置、速度、角速度、角度の 6 変数のうちのいずれか N_r 個である。このとき、モジュールの数は ${}_6C_{N_r}$ であり、$N_r = 6$ のときモジュールは一個となる。また、船はアクセルと舵の 2 種類の行為を持っている。試行錯誤を通して各モジュールは環境や他のモジュールとの関係に適応する。

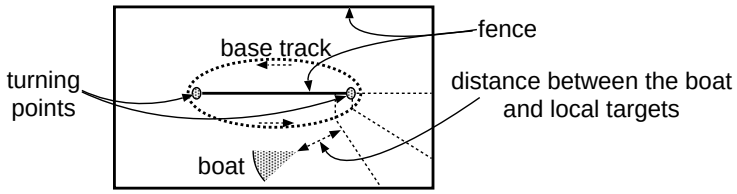

図 2: 船の操舵問題

4.2 実験結果

図 3 に、モジュールの学習する変数関係の数によって学習速度がどのように変化するかを示す。縦軸は 1 周にかかるステップ数、横軸はステップ数である。この図より、変数 6 つをすべてを持つ単一モジュールの場合より、変数の数が少ない複数モジュールの場合（$N_r = 3, 4, 5$ の場合）の方が学習速度が高速であることが分かる。したがって、モジュール化によって学習が効率化される状況もあることが分かる。

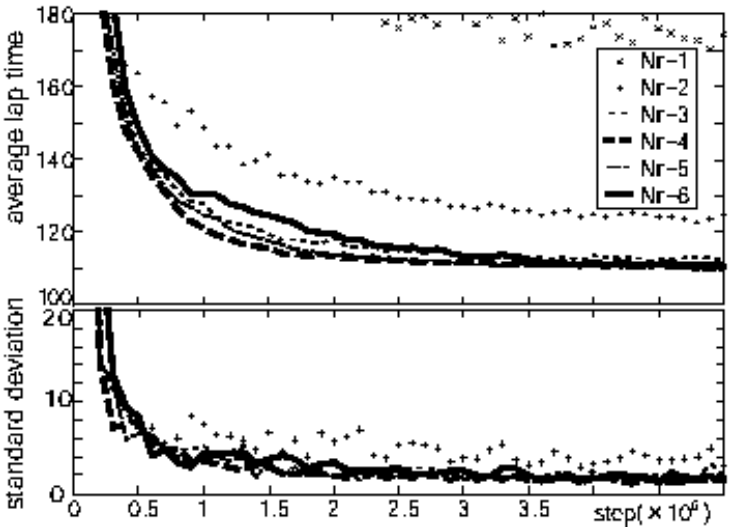

図 3: 操舵性能の比較

また、船がコースを周回している際にどのモジュールがもっとも行為選択に寄与しているかを図 4 に示す。図中の点線は船の軌跡、点線脇の数字は、もっとも活性化しているモジュールの番号を表す。この図から、船の状況に応じて活性化するモジュールが変化している様子が分かる。これは、船がある状況にあるときどのモジュールが活性化するかという関係のネットワークが形成されており、一つの系を作っていると言える。

つぎに、船のセンサを一つ故障させた場合、活性化するモジュールがどのように変化するかを図 5 に示す。この図より、前図 4 にはあまり見られなかったモジュール 6,9,17 が活性化する頻度が高くなっていることが言える。つまり、センサ故障という環境変化に対して、モジュール間の関係が変化し、活性化するモジュールが変化することで適応していることが分かる。

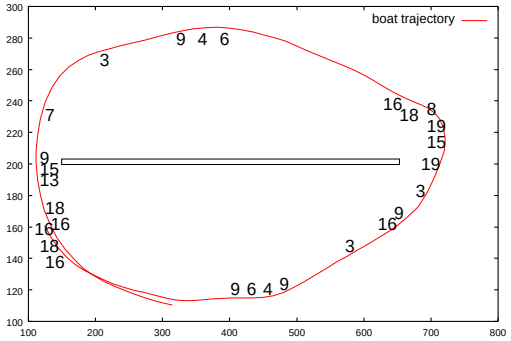

図 4: 船の状況による動的役割

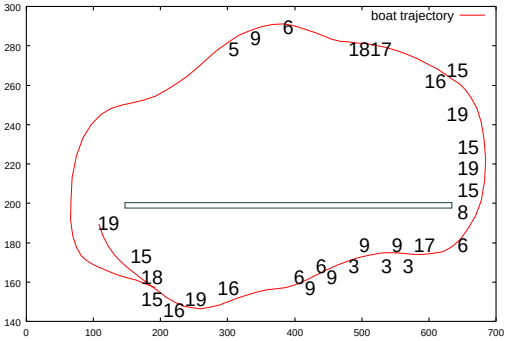

図 5: 環境の変動によって変化した役割

4.3 考察

エージェントの関係が時間を追ってどのように変化するかを、図 6, 7 に示す。図 6 は最初の一周目において、選択された行為をどのモジュールがどれほど支持したかを示しており、図 7 は 5 周目において、選択された行為をどのモジュールがどれほど支持したかを表している。横軸はそれぞれステップ数であり、横軸の最大値が異なっているのはコース一周にかかったステップ数が異なるためである。縦軸は、実行された行為に対して各モジュールが見積もった一致度と評価値の積（＝寄与度）である。

この図より、船が環境とインタラクションしている間にモジュールの間でもインタラクションが行われ、はじめ均一であったモジュール間の関係が徐々に変化していることが分かる。すなわち、個々のモジュールが相対的に個性化してゆき、船の行為選択において役割が分化していく過程が見て取れる。たとえば、図 7 において、タイムステップ 30-50 の間は船がコーナリングしているが、船の行為選択に寄与したモジュールの関係が、よく寄与するモジュール、多少関係するモジュール、ほとんど寄与しないモジュールというように分かれている。

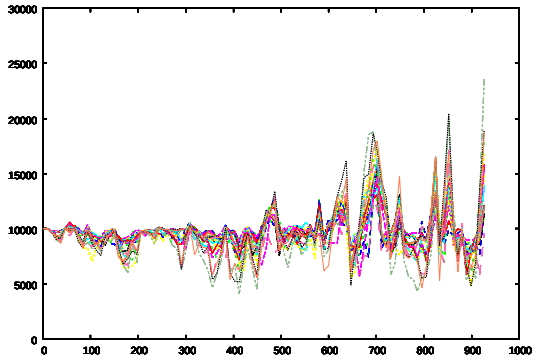

図 6: 1 周目における行為選択に対する各モジュールの寄与度の変化

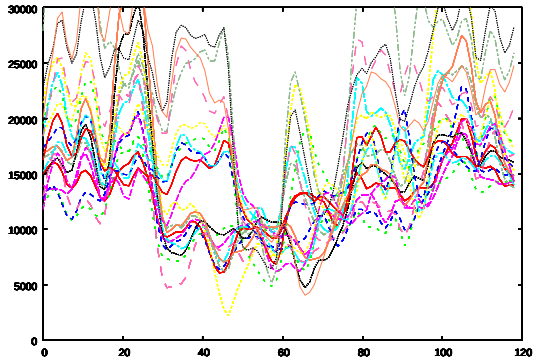

図 7: 5 周目における行為選択に対する各モジュールの寄与度の変化

言い換えると、直進する役割、コーナーを回る役割などの名前づけられる役割は一つのモジュールに担われているのではなく、モジュールの間のインタラクションによってモジュール間の結びつ

きが動的に変化することによって実現されていることがわかる。このことを図で表現すると図 8 のようになる。

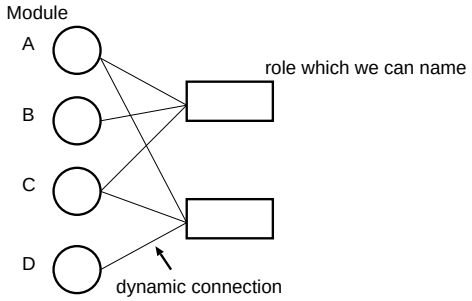

図 8: モジュールの動的な結びつきによって実現される役割

4.4　今後の課題

現在のところ、モジュールをどれだけ用意すれば学習効率が良くなるのか、どのような環境変動であれば頑健性が期待できるかについてはよく分かっていないので、今後調べていく必要がある。

5　おわりに

本稿では、未知環境や動的環境において学習空間が大きくなる問題に対して学習空間をモジュール化する場合を取り上げた。学習空間のモジュールに対してあらかじめ役割を決めておくことが困難であることを説明し、モジュールの役割がモジュール間の関係と環境の状態に応じて動的に変化することを示した。

参考文献

[1] R. S. Sutton and A. G. Barto: *Reinforcement Learning: An Introduction*, The MIT Press, 1998.

[2] 福岡正信: 無 [III] 自然農法, 春秋社, 1992.

[3] C. Watkins and P. Dayan: “Technical Note:Q-learning”, *Machine Learning*, **8**-3/4, pp.279–292, 1992.

[4] 山岸栄輝，堀内匡，川上浩司，片井修: “部分的条件省略ファジィルールを用いた強化学習”, 計測自動制御学会論文集, **37**-12, pp.1178–1185, 2001.