

外部報酬に導く内発的報酬の生成機構

A Modular Intrinsic Rewards Generation System

竹内誉羽 庄野修 辻野広司
 Johane Takeuchi Osamu Shouno Hiroshi Tsujino

株式会社ホンダ・リサーチ・インスティテュート・ジャパン
 Honda Research Institute Japan, Co., Ltd.

Inspired by intrinsic motivation that is thought to play a crucial role in animal development and learning, several artificial learning systems with built in intrinsic rewards were recently studied. Barto, Singh and Chentanez suggested intrinsically motivated reinforcement learning in which novelty based intrinsic rewards guided learning to earn external rewards. In their implementation, intrinsic rewards were generated when predetermined events occurred. Their algorithm took place only in deterministic transitions. We suggest intrinsic rewarded learning without these restrictions. The system consisted of neural networks equipped with a modular reinforcement learning algorithm suggested by Doya *et al.* The modular system that decomposes the observed state space stabilized the intrinsic rewards calculated from prediction errors in probabilistic environments.

1. はじめに

内発的動機付けによる学習は動物の発達学習において重要であると考えられている。発達ロボティクスなどの研究では、環境との相互作用から内発的動機づけを作り出し、学習すべき事柄をロボットが自ら決める (e.g. [Oudeyer 07])。これらの研究では、自らスキルを獲得していくことで、人や動物と同じような行動パターンがロボットにも自律的に発現することを期待している。しかし、将来の人型ロボットや会話エージェント等への応用を考えた場合、発達の手法だけでなく、やはりタスクを遂行するための学習機構が必要になるであろう。また発達ロボティクスとは別に、Bartoらが行動系列を学習させるために内発的動機付けを用いた強化学習 (intrinsically motivated reinforcement learning: IMRL) を提案している [Barto 04]。彼らは外部報酬に至る行動系列の学習が、内発的報酬によって加速されることを示した。しかし、その実装においては内発的報酬が生成される条件を、あらかじめ設計者が定めなければならず、また決定論的な状態遷移にのみ対応している。その他の内発的報酬と強化学習を使った多くの研究では、環境のモデル構築や探索のために内発的報酬が使われているものが多い。

現在の機械学習では、設計者が細心の注意を払って設計し、定式化された個々の問題を学習可能とする。しかし人や動物はたとえ定式化されていなくとも、学習する柔軟な能力を持ち、それによって実環境に適応している。これを機械で実現することは非常に困難なことであり、発達ロボティクスの究極の目標もここにあると思われる。IMRLの研究でも、そもそも機械学習に動物の学習のような柔軟性をもたせるために内発的報酬を導入したのであるが、実装においてはそれが十分に実現されていない。我々は、IMRLの研究で示された「内部報酬を使った外部報酬に至る行動系列の学習」を出発点として、実環境に適応する学習システムの実現に取り組んでいる。

ここでは内発的報酬を遷移の新奇性に応じて生成する仕組みの導入により、生成される条件をあらかじめ設定しておく必要性を除いた。この新規性の判定のために、ニューラルネットワークによる状態予測を用いた。その際、確率的な遷移に対応

するために、Doya等のModule強化学習 [Doya 02] を適用した。これらによって、以下のことが可能になった。

- ニューラルネットワークの実装による「内部報酬を使った外部報酬に至る行動系列学習」の実現。
- 新奇性に応じた自動的な内発的報酬の生成。
- 確率的な遷移に対応可能。

2. 学習モデル

2.1 単一モデル

我々は、[Takeuchi 06]においてニューラルネットワークを用いたモデルを構築した。観測状態を入力情報とし、行動ごとの状態行動価値関数値 (Q 値) を出力する二層のニューラルネットと共に、直前の観測状態と行動選択の状態から状態予測をするための三層のリカレントニューラルネットワークを使った。状態予測の学習のためには通常非線形な逆伝播法を使った。この状態予測の誤差により、状態遷移の新奇性に応じた内発的報酬が作られる。以上を単一モデルと呼ぶ。しかし以上の単一モデルでは確率的な遷移には対応できなかった。確率的な遷移の場合、状態予測が平均値に収束してしまうため、適切な内発的報酬を作り出せないことが原因であった。

2.2 モジュールモデル

そこでモジュール強化学習を応用した学習モデル (モジュールモデル: 図 1) を構築した。確率的な状態遷移を個々のモジュールに振り分けることで、問題の解決を試みた。基本的なモジュール化の方法はDoyaらのモジュール強化学習 [Doya 02]と同様である。すなわちモジュールごとに状態予測を行い、その予測と実際に観測された状態との一致度に応じてモジュールを選択し、強化学習を行う。個々のモジュールは、基本的には単一モデルと同等である。またモジュールとは別に、共通した履歴情報をつくるためのリカレントニューラルネットワーク層 (ESN) を用意した。

3. 結果

ここでは得られている結果の一部を示す。テストのために、図 2a に示される状態遷移を用いた。これは単純な Markov 決定過

連絡先: 竹内誉羽、株式会社ホンダ・リサーチ・インスティテュート・ジャパン、〒351-0188 埼玉県和光市本町 8 - 1、Tel:048-462-5219、e-mail:johane.takeuchi@jp.honda-ri.com

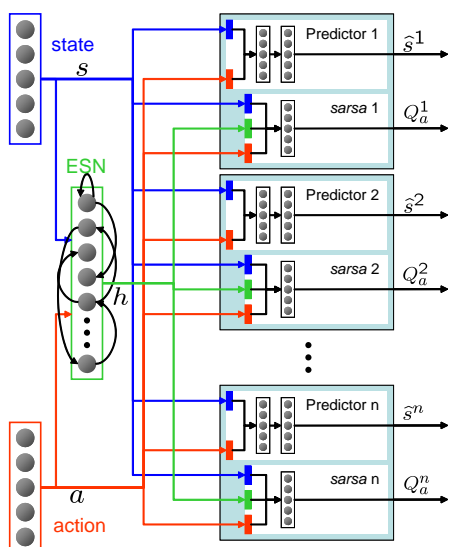


図 1: モジュール強化学習を行うニューラルネットワークの模式図

程 (MDP) であり、 $s_0 \rightarrow s_1$ の遷移だけが確率的な遷移になり、他は常に決定論的である。また、状態 s_5 のときに、行動 a_5 を選んだ時のみ外部報酬 $r^{ex} = 1$ が与えられる。すなわち、状態 s_0 から始まって、行動系列 $a_0 \rightarrow a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow a_4 \rightarrow a_5$ を選択したときのみ、外部報酬が与えられ、それ以外の遷移はすべてゼロである。

決定論的な遷移の時 (図 2b) は、単一モデルでも内発的報酬があれば、行動系列の学習ができる。また内発的報酬がない場合はどちらのモデルでも、学習速度がかなり遅くなってしまふ。

さらに遷移の一部を確率的状態遷移に変えると、単一モデルではもはや学習不能になる (図 2c)。しかし内部報酬を使ったモジュールモデルでは、きちんと学習が収束する。この場合、一つの遷移が $p = 0.3$ の確率的な遷移で他の五つが決定論的な遷移であるので、平均値が $1/0.3 + 5$ に収束するはずであるが、この結果はこれに合致している。

4. おわりに

ここで使った状態遷移のように、シーケンスの最後にならなければ報酬がないような問題を on-line 学習で扱うには内発的報酬が必要になると考える。ここでは、状態予測誤差を使った内発的報酬の生成にモジュール強化学習が有効であることをひとつの例を使って示した。今後は、このアイデアの有効範囲の明確化と拡張、ならびに具体的な応用を目指したいと考えている。

参考文献

[Barto 04] Barto, A. G., Singh, S., and Chentanez, N.: Intrinsically Motivated Learning of Hierarchical Collection of Skills, in *Proceedings of the 3rd International Conference on Developmental Learning (ICDL)* (2004)

[Doya 02] Doya, K., Samejima, K., Katagiri, K.-i., and Kawato, M.: Multiple Model-Based Reinforcement Learning, *Neural Computation*, Vol. 14, pp. 1347-1369 (2002)

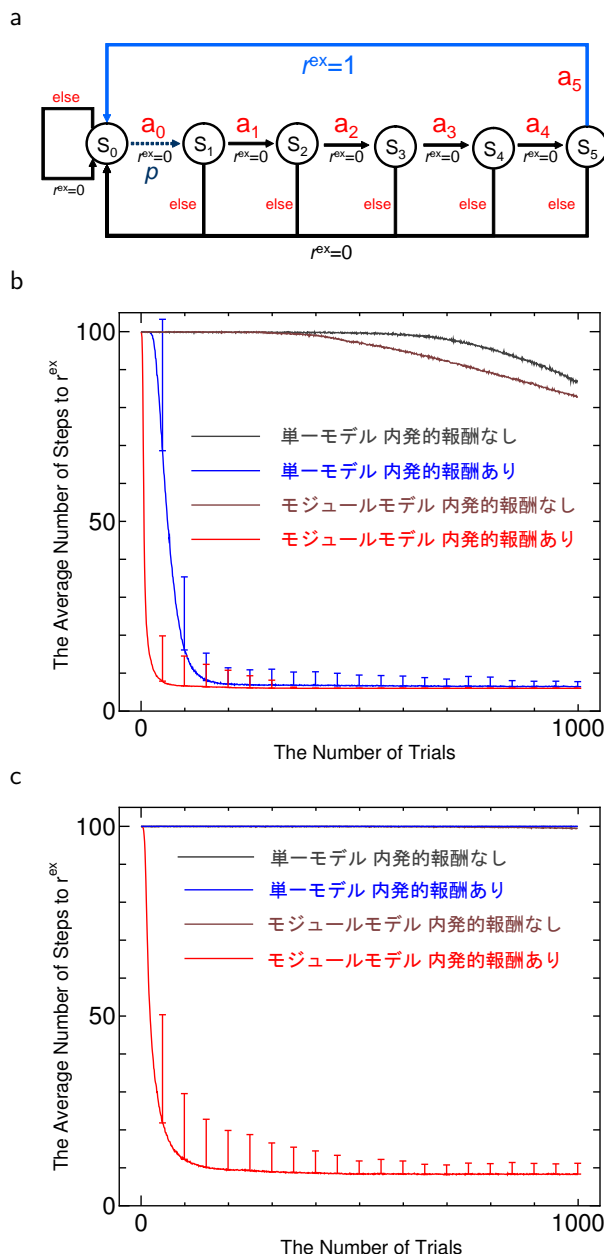


図 2: MDP 環境でのテスト結果. (a) 使用した状態遷移. 外部報酬は $r^{ex} = 1$ (b) 決定論的な状態遷移 $p = 1.0$ での学習曲線 (2000回平均) (c) 確率的な状態遷移 $p = 0.3$ での学習曲線 (2000回平均)

[Oudeyer 07] Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V.: Intrinsic Motivation Systems for Autonomous Mental Development, *IEEE Transactions on Evolutionary Computation*, Vol. 11, No. 1, pp. 265-286 (2007)

[Takeuchi 06] Takeuchi, J., Shouno, O., and Tsujino, H.: Connectionist Reinforcement Learning with Cursory Intrinsic Motivations and Linear Dependencies to Multiple Representation, in *Proceedings of 2006 International Joint Conference on Neural Networks (IJCNN'06)*, pp. 54-61 (2006)