3A2-4

環境との相互作用に基づく法則性抽象化の機構
A Mechanism of Rule Abstraction Through Interaction With Environment

赤田 庸平[*1] Yohei Akada
牧野 貴樹[*2] Takaki Makino
高木 利久[*1*2*3] Toshihisa Takagi

[*1] 東京大学大学院新領域創成科学研究科 Graduate School of Frontier Sciences, the University of Tokyo
[*2] 東京大学総括プロジェクト機構 Division of Project Coordination
[*3] 情報・システム研究機構ライフサイエンス統合データベースセンター Database Center for Life Science, Research Organization of Information and Systems

A higher organism is able to grasp and reorganize not only the surrounding environment through sequential interaction but a rule underlying it. This ability is one of the important components that provide the flexible adaptability of the organism. In this research, we propose a neural-network based reinforcement learning model that learns to choose best actions based on abstracted state transition rules, which is also learned from interaction.

1. はじめに

ヒトを始めとする高等生物は，変化し続ける環境の中で適応的に行動し続けることができる。適応的な行動選択ができるのは，適応的に行動学習ができているからであり，この仕組みを理解することは重要である。このような研究に有効な計算理論として強化学習がある。強化学習は，学習主体（Agent）が環境と相互作用するなかで，状態遷移に伴って得られる報酬を間接的な教師として総報酬を最大化するように適応を行う機械学習の枠組であり，不確実性のある環境において目先の報酬のみに囚われない行動系列を学習できる点を特長とする。代表的な強化学習である Q 学習では，状態空間に対し価値関数を割り当て，これを適切に更新することにより総報酬の最大化を実現する。

しかしながら，環境が変化し続ける場合，総報酬最大化のための行動系列は，環境が変化するたびに最初から学習する必要があるため，適応的に行動選択できるようになるまで非常に時間がかかる。もし，環境が変化しても，変化前の環境（多経験環境）で学習した行動系列を，変化後の環境（初(少)経験環境）での行動において再利用することができるならば，環境が変化するたびに最初から学習する必要がないため，適応的な行動系列をより迅速に学習できることが考えられる。一方，行動が多経験環境での経験に強く依存するという性質は，常に，新規の環境に対する適応性を損なうという問題点をはらんでいる。そこで本研究では，新規の環境に対する探索を促進しつつ，環境に潜在する法則性を学習することで，多経験環境での学習結果を初(少)経験環境において再利用できる強化学習モデルを提案することを目的とする。

学習において探索を促進する関連研究の一つに，内発的動機付けによる強化学習がすでに提案されている [Barto 2004]。この手法では，少経験の状態変化に対応する特別な状態に対して内的報酬(環境から与えられる報酬とは独立した学習者内部の仮想的な報酬)を付与することで探索を促進している。しかしながら，どの状態が特別な状態であるかは研究者によって決められているという点で，その探索の促進手法は課題設定依存的であると言える。これに対して，ニューラルネット（NN）の学習誤差を内的報酬として用いる試みがなされているが，検証に用いられている課題が状態空間は 1 次元で全状態数は 6 個と解かせている課題が簡素であること，NN が行なっている学習は状態予測であり課題の変容に対する堅牢性がないという問題点が存在する[Takeuchi 2006]。課題の本質を熟知した研究者からの教示に依らずに探索を効率的に促進するためには，課題設定に依存しないような内的報酬を与える指針が必要である。

連絡先：赤田庸平，東京大学大学院新領域創成科学研究科，〒277-8568 千葉県柏市柏の葉 5-1-5，04-7136-3982，akada@cb.k.u-tokyo.ac.jp

2. 提案手法

提案手法では，①状態遷移規則（遷移規則）を学習し（図 1），その結果を状態の一部として用い，②遷移規則の学習誤差（これは少経験の遷移規則に対応する）を内的報酬として利用する（図 2）。

①では，課題に潜在する法則性の本質は遷移規則の共通性と価値関数の共通性が相関する点にあるという着想に基づき，遷移規則で状態をグループ化することにより，適応的な行動学習のための指針を与えている。これは，共通の遷移規則を持つ状態は類似する価値関数を持つことが予想されることに基づいている。これにより，或る多経験状態と同じ遷移規則を持つ状態は，たとえ少経験状態であっても汎化により価値関数の学習が進み，また，環境が変化しても，遷移規則が同じである状態に関しては，学習済みの価値関数の情報が利用されるため，最初から学習する必要がないことが期待される。

また，探索と遷移規則に基づく状態のグループ化を首尾よく促進するには，多経験の遷移規則を経験することよりも少経験の遷移規則を経験することがより優先されるような行動学習のための指針が存在していることが要求される。本手法では，②の導入により（これは少経験の遷移規則に内的報酬を与えることに相当する）この要求を達成する。本手法ではこの指針として遷

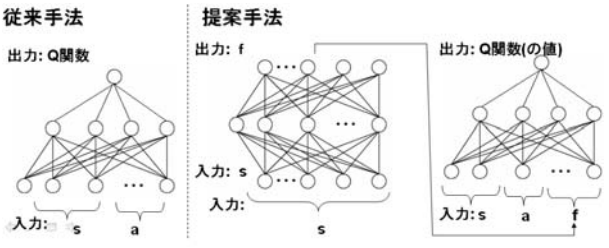

図 1　提案手法①

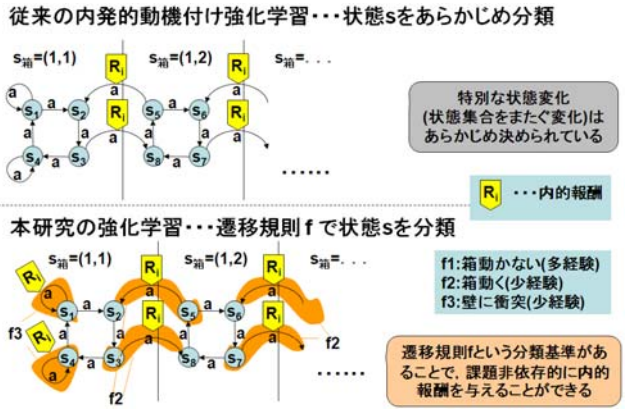

図 2　提案手法②

移規則の経験頻度を用いる。これは課題設定に依存しないため，Agent は何が内的報酬であるかをより自律的に発見していると言える。提案手法と従来手法の違いを図 2 に示す。両手法とも，状態集合間をまたぐ状態変化に対して内的報酬を与える点は共通している。従来手法においては，状態集合はあらかじめ与えられるが，提案手法においては，状態集合はあらかじめ与えられず，学習によって初めて獲得される。結果，内的報酬の大きさを，状態集合の学習(写像 f の学習)の進行に伴ってその都度決定しつつ，汎化状況を自律的に改善することが期待される。

3. 実験と考察

提案手法の有効性の検討には Goal 変化箱押し課題を用いた。この課題は，許容ステップ数以内に Agent が箱をゴールまで押し進めることを一つのエピソードとするエピソード課題である(図 3)。Agent は自身，箱，Goal の位置を知覚することができるが，それ以上の情報は，報酬以外には何も与えられない。また，フィールドの範囲外に箱を押し出すことはできない。20 ステップ数以内に課題を達成できれば「成功」，できなければ「失敗」として扱い，そのエピソードは終了する。エピソードが終了し新たにエピソードが始まるたびに Agent ，箱，ゴールの配置はランダムに初期化される。Agent は，配置が変化するたびに，ほとんど毎回「似てはいるが厳密には異なる」課題に取り組むこととなる。

1000 エピソードごとの成功率で評価を行なった。各 NN は，過去 8 ステップ分の入出力情報に対してトレーニングを行なった。また，方策として epsilon-greedy を用いた。その結果を図 4 に示す。control(両手法とも用いない場合)に比べ，①(遷移規則の学習結果を状態の一部として用いる場合)，②(少経験の遷移規則に内的報酬を与えた場合)のほうが早く学習しており，また，①＋②(両手法を併用する場合)はさらに学習が早くなっていることが分かる。また，図 3 の元になった各試行の学習曲線を調べると，①および①＋②では，学習曲線の踊り場部分が

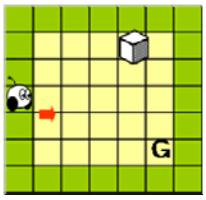

図 3　Goal 変化箱押し課題

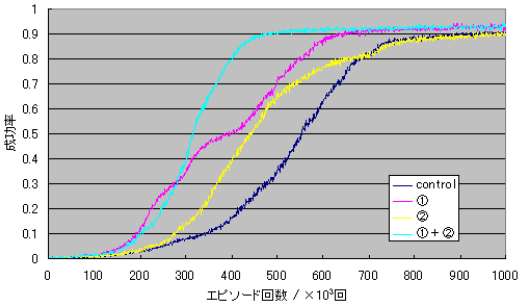

図 4　成功率(20 試行平均)

小さくなっていること，また，②および①＋②では，学習の開始と終了が早期に起こっていることが確認された。

①の導入によって，control や②で起こっていた不適切な汎化が回避された結果，踊り場部分が小さくなっているものと考えられる。②の導入が学習の開始と終了を早める効果があることについては，遷移規則の経験頻度に応じて内的報酬を与える手法によって探索が促進された結果であると考えられる。そして，①＋②においては，①と②の導入がもたらす効果が，互いに干渉せずに生じたことが，4 つの実験群のなかでの最良の成績をもたらしたものと考えられる。

4. 結論

価値の類推を伴う従来の強化学習手法は，異なる状態を同じ状態であると見なすための指針の不在，課題の変容に対する堅牢性の欠如という 2 つの困難に直面していた。本手法は，或る状態から後続する可能性のある状態への写像を課題の背景にある法則性と捉えてこれを学習し，この結果を探索の促進と状態のグループ化において利用することでこの困難を解決する。検証の結果，提案手法は，法則性学習の情報が行動に利用できるため，課題を解く過程で何が本質的に重要であるかを効率的に学習できることが示された。

これらの性能は，生物個体が，時々刻々と少しずつ変化を遂げる現実の環境に対して迅速に順応するために必要とされる汎用性の高い機能を提供する。環境に内在する法則性を学習する能力を実現する計算モデルを提案することは，これらの能力の基礎にある能力を実現する機構を明らかにすることであり，これに資する本研究は，人工知能，認知科学，神経科学の分野に貢献することが期待される。

参考文献

[Barto 2004] 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)

[Takeuchi 2006] 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)