3C1-1

自律的タスク理解とモデルベース強化学習処理の自動構成

Autonomous task understanding and automatic construction of model based reinforcement learning process

大東 優[*1] 大森 隆司[*2] 石川 悟[*3] 森川 幸治[*4]
Ohigashi Yu Omori Takashi Ishikawa Satoru Morikawa Kouji

[*1]北海道大学大学院 情報科学研究科
Graduate School of Information Science, Hokkaido Univ.

[*2]玉川大学 学術研究所
Tamagawa University Research Institute

[*3]北星学園大学 文学部
School of Humanities, Hokusei Gakuen Univ.

[*4]松下電器産業 (株) 先端技術研究所
Advanced Technology Research Laboratories, Matsushita Electric Industrial Co.,Ltd.

The traditional Reinforcement Learning(RL) supposed complex but single tasks. When the traditional RL agent faced a task similar to learned ones, the agent must re-learn the task from the beginning because of its unuse of learned result. In this paper, we propose a reinforced learning that can realize quick acquisition of actions for tasks similar to previously learned ones. Ours is a model-based RL that employs task model constructed by combining primitive local predictors for environmental dynamics prediction. To evaluate effectiveness of the proposed method, we performed a computer simulation using a simple ping pong game and variations of it, and demonstrated very quick adaptation to new tasks.

1. はじめに

現在，機械学習の分野では学習タスクの種類に応じて様々な学習手法が提案され，単一の複雑なタスクに対しては効率の良い学習が可能となってきている．しかし，それぞれの学習タスクにおいて「どの情報や知識を用いてどのような処理手順で学習するのか」という学習処理そのものの設定は，学習機械が自律的に行うわけではなく，技術者や研究者などの人間が個々の問題に応じて行い，学習機械に組み込んでいる．

また，個々の問題に対して人間が取り組んで解決することから派生して，機械学習のモデルの多くは単一のタスクを学習することを想定している．そのため，ある学習タスクで獲得した知識はそのタスクに特化していることが多く，類似した別の学習タスクに対してそれまで獲得した知識を再利用することはできず，一から再学習する必要がある．たとえ獲得した知識が再利用可能な場合でも，その知識をどのように利用するかという学習タスクの設定には人間設計者の介入が必要であり，現在の機械学習は複数の学習タスクに対しては柔軟には適応できない．

つまり，ある学習機械，あるいは学習モデルが様々な学習タスクに適用されてその問題を解決できることの大きな要因は，学習機械が学習する前段階に人間が行う学習タスクの分析と，その結果に応じた学習の処理過程の構築にある．それは，多様な問題解決にあって真に知能を持って発揮しているのは人間であって，現状の学習機械は新しい問題に自律的に対応する能力は持っていないということを意味している．

ところが将来，ロボットのような知的機械が日常生活の中で人間と共に活動する場面を考えると，ロボットは変動し続ける実世界において，全く新奇ではないが，学習タスクとしては異なる様々な新タスクに次々に遭遇する．我々はこのような環境を多重タスク環境と呼ぶ．その状況でロボットは，次々と表れる学習タスク群に対して，何をどうやって学習するかを人間の手を介さず自律的に決定し，適切な行動を学習する必要がある．そのためには，個々の学習タスクに対する学習の効率化よりも，学習タスクそのものを自律的に設定する手法の開発が必要となる．

我々は，この問題に対してモデルベース強化学習の枠組みを用いてアプローチする．モデルベース強化学習とは，環境のダイナミクスのモデルを用いて効率の良い学習を実現する強化学習の枠組みである [Sutton 1990]．環境モデルは，現在状態と選択された行動からの次の状態の予測や，ある状態においての報酬の獲得の可能性を判定する予測の機能を持つ．あるタスクの解決に必要な環境モデルを構築するという作業は，モデルベース強化学習においてタスク設定の主要な部分を実現することに相当する．

従来の手法では，環境モデルは学習エージェントが環境との相互作用に基づいて徐々に獲得するか，ヒトの手によってあらかじめ組み込まれている．しかし後者は本研究の自律的タスク設定にとっては想定の外である．そして前者の方法においては，環境モデルの獲得に時間がかかることが，従来のモデルベース強化学習の研究の主要な課題であった．しかし，新奇タスクごとに環境モデルを学習するのではなく，過去の類似タスクの経験に基づく知識の再利用によってすばやいタスク設定を自律的に行う方法は，この問題の根本的な解決につながる．

そこで本稿では，モデルベース強化学習においてタスクの表現までを含む環境モデルをタスクモデルと呼び，新奇の問題に対するタスクモデルを過去の経験から得た知識を再利用して自律的かつ動的に構築する手法を提案する．我々は，複数の因果関係が含まれる単純なテレビゲームを題材としてシミュレーションを行い，提案手法の有効性を検証する．

このような予測モデルを用いたモジュール型のネットワークとして，Mixture of experts [Jacobs 1991]，MOSAIC [Kawato 1998]，MMRL [Doya 2002] が知られている．我々が認知的な課題に対する問題解決場面を想定し，局所的な情報を入力とした予測モデルを単位としているのに対し，これらのモデルは運動制御の課題を対象とし，環境のすべての状態を入力とした予測モデルを単位としている点で我々とは異なる立場

連絡先: 大東 優，北海道大学大学院情報科学研究科，札幌市北区北 14 条西 9 丁目，011-706-6815，y_ohigashi@complex.eng.hokudai.ac.jp

である．

2. 局所的な予測モデルの組み合わせによるタスクモデルの動的構築

我々が提案するタスクモデルは，環境に存在する局所的な因果関係の組み合わせにより構築され，現在状態と行動から報酬が得られるまでのタスク状態の変化過程の予測モデルである．つまり，タスクモデルはタスクのシミュレーターとして機能する．

タスクモデルの自律的な構築のためには，我々は現在のタスクがどのような因果関係で成り立ち，どういう状態でどういう行動を取ると報酬が獲得できるのか，といった環境に関する知識を発見して表現する必要がある． その目的のため我々は，実環境に対する以下の二つの仮説を立てる．

- タスク環境の状態を表す変数群の間には，物理法則や因果関係などによる局所的な制約がある．
- 類似のタスク群はお互いにそれらの制約を共有している．

つまり，タスク環境は局所的な物理法則，報酬獲得の因果関係，そして環境とエージェントとの関係によって表現されており，さらに個々のタスクは，複数の因果関係を含んでいると仮定する．

このような実環境に対して，学習タスクの自律的な設定や多様なタスクへの即時適応を実現するためには，そのシステムは以下の二つの機能が必要となると我々は考える．

1. タスクの各場面において働いている局所的な因果関係の発見と，変数間の予測の学習，
2. 変数間の局所的予測モデルの動的な組み合わせによる，タスク全体の予測モデル (タスクモデル) の構築．

そのために我々は第一に，学習タスクに存在する局所的な因果関係を変数値の変化を予測する局所的な予測モデルによって表現した．つまり，環境に含まれる因果関係は変数間の局所的な予測モデルとして学習し獲得され，その環境についての知識として蓄積される．それらの知識は，環境の自動的な状態変化を予測するモデル，報酬の発生を予測するモデル，行動に対する状態変化を予測するモデルの 3 つに分割する (図 1)．それらの知識は，それがどのような条件の時に環境を予測可能かという適用条件 (selector) とセットになって学習される．獲得した予測モデルを用いて予測を行うことで，現在の学習タスクに対してどの部分が予測可能でありどの部分が予測不可能であるかが判断できる．予測不可能部分についてその部分に対する予測モデルを新たに生成して学習することで，類似タスクに移行した際の変更部分においても知識獲得が実現できる．

個々の予測モデルの適用条件 (selector) が獲得された後は，観察した環境の状態に対して有効な予測モデルが選択可能となる．すなわち，環境からの入力を selector 群に与えるとその瞬間の状態に対して適用可能な selector 群が高い出力を出し，それらの間で最大の出力をしたものに対応する予測器を使用し，次の時刻の環境状態を予測する．その予測のサイクルを次々と適用することで，多様な状態に対して適切な予測を続けることができ，現在状態から比較的長い時間にわたって環境の変化を予測することができる (図 2). すなわち我々は，環境の観察からボトムアップでその局所予測モデルを選択して組み合わせ，結果として環境のうちでタスクに関わる部分の予測を動的に行うタスクモデルの自律的な構築を実現する．

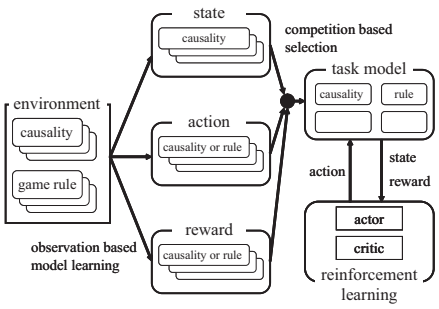

図 1: 環境の局所予測モデルの動的な組み合わせによるタスクモデルの構築

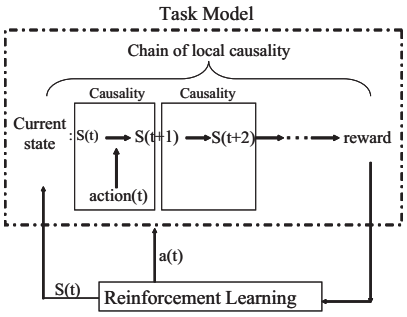

図 2: 強化学習とタスクモデル間の処理の流れ

類似タスクに移行した際においても，以前のタスクで獲得した局所予測モデルが使用される．それが可能なのは，局所予測モデルが環境中の全変数ではなく，その一部の間にある関係のみを利用して予測を行うためである．新奇のタスクの状態空間が以前のタスクと異なっても，注目する現象に関わる部分変数がおなじであれば，それらの間の予測器と selector は利用可能である．局所予測モデルが使用する変数が限定されているほど，そのモデルの適用範囲は広くなる．そのため，新奇タスクで不足している知識を予測モデルとして追加的に獲得した後は，それらと既存の知識を観察に基づきボトムアップ的に組み合わせるだけで，新奇タスクに対するタスクモデルは素早く構築でき，多様なタスクへの柔軟な適応を実現する．

タスクモデルが構築された後は，我々が以前提案した PRLmodel と同様に [Ohigashi 2003]，タスクモデルを利用してメンタルなシミュレーションを行い，現在状態と予測された状態に対して強化学習を用いて最適な行動を学習する．

3. 提案モデルの学習アルゴリズム

3.1 局所予測モデルの獲得

局所的な予測モデルは，環境中のある事物の現時刻 t における状態 s_t から，次時刻 $t+1$ の状態 s_{t+1} を予測するモジュールである．ここで，各予測モデルの入力される状態 s_t は観測できるすべての状態変数群ではなく，そのイベントにかかわる部分的な状態変数群である．さらに，個々の局所予測モデルはある事物がかかわる全ての現象の予測をする訳ではない．ある局所予測モデルはある事物のある瞬間，あるいはある因果に基づく動作 (例えば慣性の法則に基づくボールの直進) を予測するように学習し，他の法則（例えば壁でのボールの反射）は他の局所予測モデルが分担する．したがって，個々の局所予測モ

デルは局所的な法則性に特化し，それらの集合で事物の状態遷移を予測する．

局所予測モデルは，学習開始時には，環境の状態変化を予測するモデル，報酬を予測するモデル，行動に対する状態変化を予測するモデルの 3 種類の予測モデルをそれぞれ一つだけ持つ．個々の予測モデルは，Predictor P^i とそれに対応する Selector S^i のペアから構成される．このとき，P^i と S^i は，以下の手順にしたがって学習される．

1. 獲得されたすべての局所予測モデルが現在状態 s_t から次の状態 s_{t+1} を予測する．ここで，i は局所予測モデルのインデックスである．$P^i(s_t) = s^i_{t+1}$
2. 次時刻 $t+1$ における状態 s_{t+1} を使って，各局所予測モデルの予測誤差 $error^i_t$ を計算する．$error^i_t = s_{t+1} - s^i_{t+1}$
3. 最小の予測誤差を持つ局所予測モデルを選択する．$winner = arg\min_i error^i_t$
4. もし，勝者の局所モデルの予測誤差が学習したにもかかわらず減少しない場合は，新しい局所予測モデルがその状態遷移を学習するために追加される．そうでない場合は，P^{winner} が s_{t+1} を教師信号として学習し，S^{winner} は 1，それ以外の Selector は 0 を教師信号として学習する．
   $if \quad error^{winner}_t - error^{winner}_{t-1} > threshold \quad then$
   add new LPM_i
   $else$ Predictor and Selector learns.
5. 1. へ戻る．

予測モデルの追加と学習を繰り返すことで，ある局所変数間の関係に特化した予測モデル群ができる．複数の予測モデルが生成された場合は，すべての予測モデルが入力に対して予測を行なって競合を行ない，勝者が予測する．勝者の予測モデルの予測誤差が上記の基準にマッチしない場合は，それが学習・予測を行い，マッチした場合は予測モデルを追加する．

Selector S^i は，入力された状態においてペアとなっている Predictor P^i が正しく予測できるかどうかを学習する．具体的には，上述の予測モデルの競合の勝者であれば 1 を，そうでなければ 0 を出力するように学習する．このとき S^i への入力変数は，必ずしも P^i と同じである必要はなく，予測が容易になる入力変数をタスク全体の変数群から選択する．この入力変数の選択は予測モデルの再利用性に大きな影響を与える．個別のタスクに依存した特殊性の高い入力変数を選択すれば再利用性は低くなり，逆にタスク非依存の一般的な入力変数であれば高くなる．後に示すシミュレーションでは，P^i と S^i の入出力変数はあらかじめ人間が指定して学習させた．しかし，個々の局所予測モデルに関わる変数の，タスク全体の変数集合からの選択は知覚情報からの因果関係の発見という重要な問題の現れであり，今後の課題となる．

3.2 タスクモデルの動的構築

環境の自動的な状態変化を予測するモデルを e，報酬を予測するモデルを r，行動に対する状態変化を予測するモデルを a とする．報酬を予測する Predictor を P_r と表記するとき，タスクモデルは以下の手順で構築する．

1. 行動 a をランダムに決定する．
2. 行動に対する状態変化の予測モデルをすべての行動 a に対応する selectors の競合により選択し，行動の結果を予測する．
   $win_a = arg\max_i S^i_a(s_t), \quad P^{win_a}_a(s_t, a) = s_{t+1}$
3. その瞬間の環境の状態変化の予測モデルをすべての環境予測モデル e に対応する S^i の競合により選択し，環境の自動的な変化を予測する．
   $win_e = arg\max_i S^i_e(s_t), \quad P^{win_e}_e(s_t) = s_{t+1}$
4. 予測された状態を報酬の予測モデルに入力する．ある報酬の獲得が予測された場合は予測状態 s_{t+1} を目標状態 $\hat{s}$ とし，それ以外であれば 3. に戻る．また，ある一定ステップ数の予測を繰り返しても報酬の獲得が予測されない場合は，予測は出力しない．
   $win_r = arg\max_i S^i_r(s_t)$
   $if \quad P^{win_r}_r(s_{t+1}) \neq 0 \quad then \quad s_{t+1} = \hat{s}$
   $else$ (3) に戻る

3.3 モデルベース強化学習

構築したタスクモデルを用いて，タスクの状態変化の予測を行う．タスクモデルが出力した予測状態を現在状態に対する目標状態とし，タスクモデルが出力した予測報酬を用いて現在状態と目標状態に対する最適な行動を TD 学習を用いて学習する．TD 誤差の式は以下のようになり，

$$TDerror = reward + \gamma * V(\hat{s}) - V(s_t)$$

この値に基づき状態価値と行動の選択確率を更新していく．タスクモデルが予測を出力しない場合は，目標状態は従来の TD 学習と同様に s_{t+1} を用いる．

この学習方式，は我々が以前に提案した PRLmodel[Ohigashi 2003] と同様の手法である．学習結果を用いることで，現在状態と報酬が獲得できるであろう予測状態を考慮した行動決定が実現でき，従来の actor-critic 手法よりも速い学習が実現できる．

4. 計算機シミュレーション

4.1 実験設定

エージェントは，図 3 に示すボールをパドルで打ち返す単純なゲームにおけるパドルの操作 (右に動く，左に動く，その場に留まる) を学習する．

各予測モデルの P^i と S^i には，3 層の NN を用いた．環境の状態変化の予測モデルの P^i_e は，ボールの x,y 方向のそれぞれの速さ $(v_x, v_y)_t$ を入力とし，1 ステップ後のボールの速さ $(v_x, v_y)_t$ を予測する．報酬の予測モデルの P^i_r は，センサー値を入力とし，報酬 (1or0) を予測する．行動に対する状態変化の予測モデルの P^i_a は，パドルの状態と選択した行動を入力とし，1 ステップ後のパドルの状態を予測する．環境の状態変化の予測モデルの S^i_e と報酬の予測モデルの S^i_r は，センサー値を入力とし，P^i_e, P^i_r が利用可能かどうか (1or0) を学習し予測する．S^i_a は，action を入力とし，P^i_a が利用可能かどうか (1or0) を学習し予測する．タスクモデル構築後の予測の反復は，100 ステップ先の予測を上限とし，それを超えた場合は，予測に基づいた学習は適用しない．

各変形ゲームに対するタスクモデルの構築は，すべて白紙の状態から学習するわけではなく，それぞれ事前知識を与えている．two ball task では，2 個のボールに同一の環境の局所予測モデルを適用すること，2 個のボールのうちより速く盤面下部に到達すると思われるボールとして，盤面の y 座標値が小さいものを選択して予測する、という処理を与えた．また，player が局所予測モデルを利用してメンタルなシミュレーションを実行するために，player に壁がどこにあるのかという情

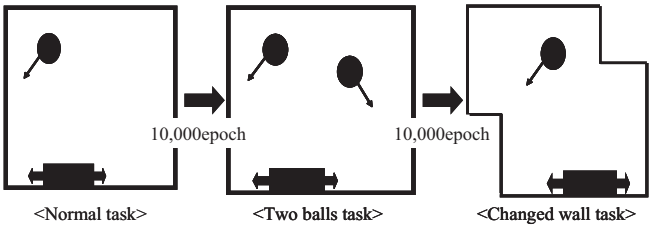

図 3: Task flow : 3 種類の類似タスクを順に学習する.

報を事前に与え，環境の状態変化の予測モデルの Selector によるモデル選択を可能にしている.

タスクモデルによって学習する強化学習部分の状態価値関数は，現在のパドルの位置とタスクモデルによって決定されたパドルの目標状態をそれぞれ 10 分割し，100 個の状態を持つルックアップテーブルにより作成した．それぞれの状態に対して，右に動く，左に動く，その場に留まるの 3 種類の行動が割り当てられ，その状態価値を学習させた．行動決定には，ソフトマックス行動選択手法を用いた.

4.2 手順

最初の 10000epoch は，normal task をプレイさせ，局所的な予測モデルを学習させる．その後，two balls task を 10000epoch，wall change task を 10000epoch と順番に学習し (図 3)，各新奇タスクでのタスクモデル再構築過程を観察した．タスクの変化時には，RL の状態価値関数はリセットした．そのときの 100epoch 毎の打ち返し率の遷移と，タスク変更後のタスクモデルの構築の振る舞いより，本稿の提案手法を評価した．ここでボールが上向きに動きはじめてから，パドルがボールを打ち返すか落とすまでを 1*epoch* と定義する.

4.3 結果

図 4 は，横軸に学習回数，縦軸に 100epoch 毎のボールの打ち返し成功率を表している (5 試行の平均)．normal task 学習時は，初期の段階では局所予測モデル群を学習する必要があるため，タスクモデルを利用した効率の良い学習が実現できない．しかし予測モデルが学習された後は，一気に行動学習が進む．10000,20000epoch のタスク変更のタイミングでは行動テーブルが初期化されて行動の成功率が下がるが，normal task の学習初期に獲得した予測モデルを再利用し，タスクモデルをすばやく構築することで，モデルベース強化学習をすばやく実現し，成功率を急激に上昇させることに成功している．normal task から two balls task に遷移する際には，normal task で獲得できないボール同士の衝突に対応する局所予測モデルが獲得されている．これにより，以前獲得した知識を再利用し現在のタスクにおいて足りない知識をのみを学習することで，すばやい学習が実現されている．また，two balls task での成功率の上限が他のゲームに比べると低いが，これはボールのスピードと比してパドルの移動スピードがそれほど速くないため，2 個のボールがほぼ同時に盤面下部に落下してきた場合や盤面下部でボールの衝突があった場合に対応できないことによる.

5. まとめ

本稿で我々は，複数の局所的な予測モデルを動的に組み合わせることで，環境のダイナミクスを予測するタスクモデルを構築する手法を提案した．この手法を評価するために我々は，タスクモデルと強化学習を組み合わせたモデルベース強化学習を用いて，類似のタスク群に対するシミュレーションを行い，各学習タスクに対する自律的な処理手順の構築と，類似のタスク群に対するすばやい適応を実現した.

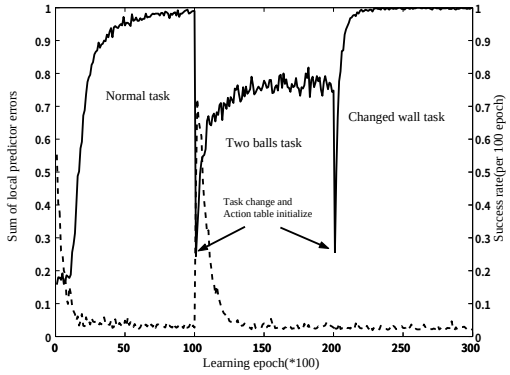

図 4: 100epoch 毎の打ち返し成功率の変化

本研究での提案手法の一つの特徴は，過去に獲得した知識を現在の問題に適用し，タスクモデルをオンラインで自律的に構築する点にある．これは，人間の知的な行動決定の一つの特徴である．知識の再利用や局所的な予測モデルの利用などが検証できる行動実験パラダイムを構築し，提案モデルとヒト行動との比較よりモデルを改良していくことで，人間の行動学習・行動決定の計算論的モデル化につながることを期待している.

参考文献

[Sutton 1990] Sutton, R.S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. Proceedings of the Seventh International Conference on Machine Learning, pp.216-224.

[Jacobs 1991] R.A.Jacobs, M.I.Jordan, S.J.Nowlan, G.E.Hinton (1991). Adaptive Mixtures of Local Experts. Neural Computation, 3, 79-87

[Kawato 1998] D. M. Wolpert and M. Kawato (1998). Multiple paired forward and inverse models for motor control. Neural Networks, 11, 1317-1329.

[Doya 2002] K.Doya, K.Samejima, K.Katagiri, M.Kawato (2002). Multiple Model-Based Reinforcement Learning. Neural computation, 14, 1347-1369.

[Ohigashi 2003] Yu Ohigashi, Takashi Omori, Koji Morikawa, Natsuki Oka (2003). Acceleration of Game Learning with Prediction-based Reinforcement Learning -Toward the emergence of planning behavior-, ICANN/ICONIP 2003, LNCS 2714, pp.786-793.