4H1-1

強化学習における効率的な転移学習適用に関する一考察

A Study on the Efficient Application of Transfer Learning to Reinforcement Learning

齋藤　碧　　小林　一郎
Midori Saito　　Ichiro Kobayashi

お茶の水女子大学大学院人間文化創成科学研究科理学専攻
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Reinforcement learning is generally employed to learn the policy of agent behaviors. In reinforcement learning, if there is no policy for a target task, the agent has to search space randomly to obtain policies, therefore, the reduction of the number of searches is regarded as a big problem to be solved. In this context, the method to apply and resuse the policy learned before to a new task is employed in the case of facing a similar task to a target task. Considering this, in this study, we propose a method which efficiently finds similar tasks and transfers policies from a source task to a target task, by applying sparse coding to transfer learning.

1. はじめに

強化学習 [1] では，エージェントが環境を探索し，与えられたタスクを試行錯誤を繰り返しながら最適な行動規則を求める．しかし，強化学習はタスクの状態を逐次的に探索しながら学習を行うため，問題点として多くの学習回数を必要としてしまうことが挙げられる [3]．そこで，エージェントの学習回数の削減を目指した研究が数多くなされており，その中の転移学習では，類似したタスクで事前に学習した行動規則を，新しいタスクにも適用し再利用することにより，新しく最初から学習しなおす必要がなくなるため，学習の効率化を図っている．しかし，転移学習において環境やタスクの類似性の定義は明確にされていないため，それぞれのタスクに応じた方法で類似度を計算する必要がある [4] が，タスクの状態数やエージェントのとりうる行動数が膨大な場合には，転移させる情報量も膨大になってしまい，類似度計算が複雑になってしまう．そこで，本研究では強化学習で得られたデータをスパースコーディング [5] を用いて基底の集合 (辞書) とスパース行列 (係数行列) に分解する．そのように複雑な類似度計算にスパース性を持ち込むことにより，従来よりも単純な表現で再現性のあるタスク分類を目標とする．

2. 強化学習

強化学習 [1] は，エージェントが環境の状態の探索を繰り返すことにより，最適な行動規則を学習する手法である．具体的には以下の 1〜3 を繰返す．1. エージェントが状態を観測する．2. 現時刻での環境において，選択することのできる行動から一つ選び実行する．3. ある環境においてある行動を実行したことに，報酬もしくはペナルティを与えて評価する．また，強化学習はマルコフ決定過程 (MDPs) として定式化されており，$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$ の四つ組で表される．ここで，$\mathcal{S}$ は状態の集合，$\mathcal{A}$ は行動の集合，その遷移確率を $\mathcal{P} = Pr\{s_{t+1} = \acute{s} | s_t = s, a_t = a\}$ で表す．また，$\mathcal{R}$ は環境からエージェントへの報酬である．エージェントの意思決定は行動規則 $\pi(s, a) = Pr\{a_t = a | s_t = s\}$ によって表され，強化学習では報酬の期待値の和を最大にする行動規則 $\pi^*(s, a)$ を獲得することを目標とする．

連絡先: 齋藤　碧，お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース小林研究室，〒 112-8610　東京都文京区大塚 2-1-1，03-5978-5708，saito.midori@is.ocha.ac.jp

2.1 Q-learning

本研究では強化学習のアルゴリズムとして Q-learning[2] を採用した．Q-learning は TD 学習の一つであり，Q 値と呼ばれる行動の評価値を最大化する．Q 値の更新式を以下に示す．

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)) \quad (1)$$

ここで，$Q(s, a) = E\big[R | s_t = s, a_t = a\big]$ であり，状態 s において行動 a を選択した時の割引収益を表わす行動価値関数である．また α は学習率であり，γ は割引率を表わしている．また，本研究では，エージェントの行動選択方法として，ϵ-greedy 選択を用いた．ϵ-greedy 選択では，ϵ の確率でランダムな行動を選択し，$1 - \epsilon$ の確率で最大 Q 値を持つ行動を選択する．

3. 転移学習

転移学習では，元タスクで強化学習により得られた方策や Q 値といった知識を，類似した目標タスクで事前知識として予め転移させておくことで，少ない探索回数で学習を行うことを目標とする．しかし，元タスクと目標タスクが似ていない場合，負の転移が発生してしまう可能性がある．そこで，どの元タスクを転移させるかを判別するために，目標タスクとのタスク間類似度を測る必要がある．また，転移学習には，タスクの状態数や行動数が多い場合には，類似度計算量も相応して増えてしまうという問題点もある．このような問題を解決するために，本研究ではスパースコーディングを導入することで，情報の質の低下を抑えた，情報量の削減をする．

4. スパースコーディング

本研究では，転移学習においてタスク間の類似度を計算する際に，スパースコーディング [5] という手法により，強化学習で得られた知識をスパースな情報にすることで，計算量を軽減した類似度測定法を提案する．スパースコーディングは以下により定式化される．

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} \quad (2)$$

ここで $\boldsymbol{y}$ は入力信号 (本研究では強化学習で得られた Q 値) を示しており，$\boldsymbol{D}$ は辞書と呼ばれる基底の集合である．また，$\boldsymbol{x}$ は $\boldsymbol{y}$ を基底の線形和で表現した際のそれぞれの基底に対応する係数行列である．スパースコーディングでは，$\boldsymbol{y}$ を $\boldsymbol{D}$ と $\boldsymbol{x}$ に分解する．また，スパースコーディングの最適化式は以下で示される．

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\frac{1}{2}} \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1 \tag{3}$$

ここで，右辺の第一項は $\boldsymbol{y}$ と復元された信号 $\boldsymbol{D}\boldsymbol{x}$ の二乗和誤差最小化を示しており，第二項はスパースな $\boldsymbol{x}$ の導出の制約を意味する．λ は正則化パラメータである．式 (3) により，最適なスパースな係数行列 $\boldsymbol{x}$ が求められる．図 1 にスパースコーディングで入力信号を分解する様子を示す．本研究では，強化学習で得られた Q 値のデータを，共通した辞書で同じ方法で分解する．そうすることで得られたスパースな係数行列の比較で転移学習における類似度計算を行う．

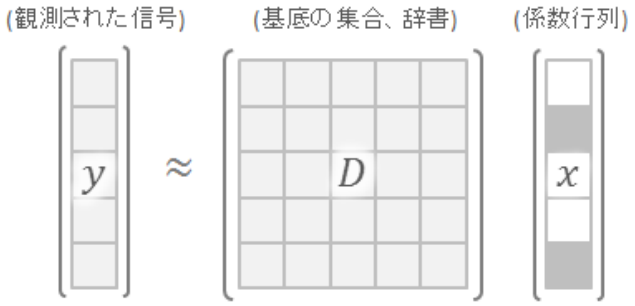

図 1: スパースコーディング

5. 実験

本実験では，20 × 30 マスの格子空間上で最短経路問題におけるタスク間の類似度を測定した．図 2 のように，スタート地点は (0, 14) であり，すべてのタスクでスタート位置は共通とした．また，ゴールの地点を変えた 1〜4 のタスクとタスク A，タスク B の計 6 種類のタスク (1:(0,0), 2:(0,19), 3:(29,19), 4:(29,0)), A:(0,2), B:(29,17)) について最短経路を求める強化学習を行った．本実験ではタスク A とタスク B を，1〜4 のどのタスクと似ているかという分類を行う．ここで，α を 0.1，γ を 0.9，ϵ-greedy 選択における ϵ の値を 0.2 とし，6 種類のタスクそれぞれで得られた Q 値をスパースコーディングの入力ベクトル $\boldsymbol{y}$ とした．スパースコーディングの手法として Lasso-lars を用いた．また，辞書は 0〜100 までの一様乱数の行列を生成し，全タスク共通の辞書とし，これを用いて，強化学習で得られた Q 値をスパースコーディングで分解し，それぞれの係数ベクトルを出力とした．スパースコーディングの入力 $\boldsymbol{y}$ の大きさを 2300 とし，出力の係数ベクトル $\boldsymbol{x}$ の大きさを 500 と設定した．ここで，スパースコーディングの入力である 6 種類のタスクの Q 値を図 3〜図 8 に示す．これらは，横軸は，20 × 30 マスにおいて選択することのできる全ての行動を 1〜2300 で表しており，また，縦軸は，それぞれの Q 値を示している．次に，図 9 に 4 種類のタスクの係数ベクトルの結果を示す．また，図 10 は 1〜4 と同様にして取得した A と B の結果である．これらも，横軸が係数ベクトル，縦軸が係数を示している．

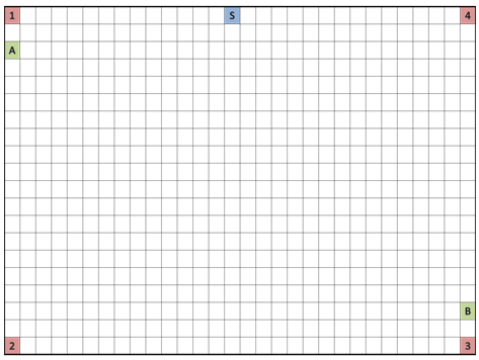

図 2: 20 × 30 マスの格子空間

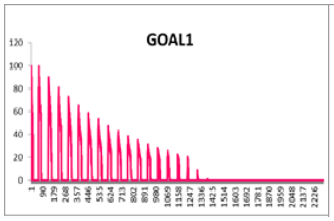

図 3: 1

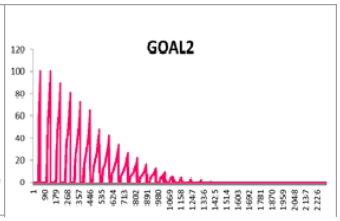

図 4: 2

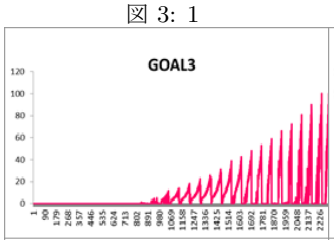

図 5: 3

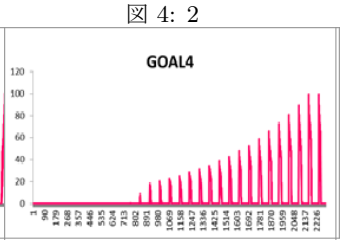

図 6: 4

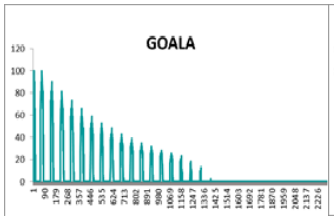

図 7: A

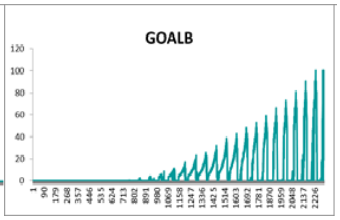

図 8: B

5.1 考察

まず，スパースコーディングにより，2300 個の入力を 500 個の出力で表現することができ，500 のうち係数が 0 でないものはごく一部となっていることが図 9 と図 10 よりわかる．よって，スパースコーディングにより情報量を削減できることを確認した．次に，表 1 に図 3〜図 8 の結果を，表 2 に図 9 と図 10 の結果を，それぞれ平均絶対値誤差で表した．これは，A と B のタスクと，1 から 4 までのタスクの誤差を示し、どれくらいの相違があるのかを数値で確認したかったため行った．表 1 より，タスク A はタスク 1 と，タスク B はタスク 3 と，一番誤差が小さく，ゴール地点の距離が一番近いものと最も相関があることがわかる．また，図 9 と図 10 では，どのタスクと類似しているか分かりづらかったが，表 2 より，表 1 と同様にタスク A は 1 と，タスク B は 3 と最も相関があることが示されている．従って，スパースコーディングによる分解は，元のデータの相関性を失うことなく，転移学習においてどの元

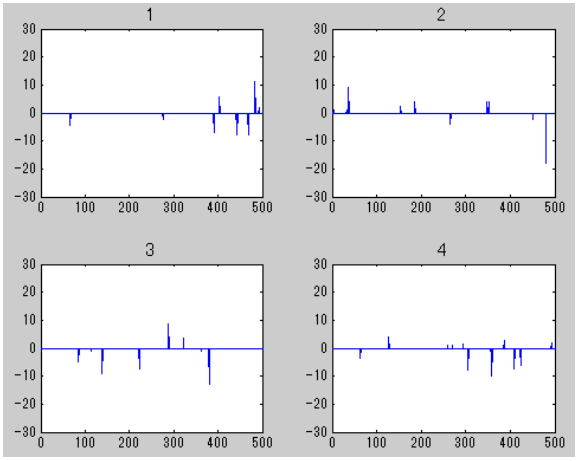

図 9: 係数ベクトル (1〜4)

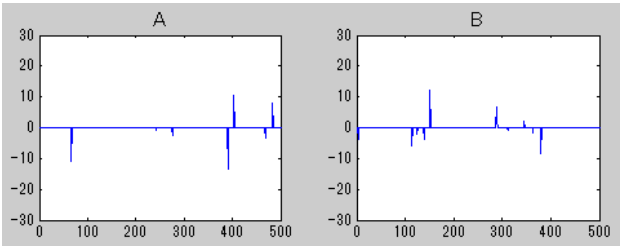

図 10: 係数ベクトル (A,B)

タスクの知識を利用するかを分類する際の類似度計算に有効であることがいえる.

表 1: 平均絶対値誤差 (Q 値)

	1	2	3	4
A	1.861872	11.84632	13.20716	10.69199
B	13.67212	14.33817	3.273166	12.84143

表 2: 平均絶対値誤差 (係数ベクトル)

	1	2	3	4
A	0.0784464	0.2	0.2	0.2
B	0.1999994	0.2	0.1148254	0.2

6. まとめ

本研究は，スパースコーディングを導入することにより，強化学習で得られた Q 値をスパースな係数ベクトルに変形し，類似度測定をする手法を提案した．これにより，計算をする対象が疎になることで，保存しておく情報量を削減することができ，またスパースコーディングで分解した後でも類似度の再現性があることを示した．今後の課題として，現在は目標タスクにおいても一度学習を行う必要があるが，オンラインによる逐次的な探索にスパースコーディングを組み込み，知識を転移させたいと考えている．また，現在は辞書に一様乱数を代入しているが，辞書学習を取り入れることで，基底のパターンを学習し，高精度な分解やデータの解析に繋げていきたいと考えている．

参考文献

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

[2] C.J.C.H.Watkins, Learning from Delayed Rewards, PhD thesis, King's College, Cambridge, UK, 1989.

[3] 高野敏明, 高瀬 治彦, 川中 普晴, 鶴岡信治, 強化学習における異目的タスク間での知識の転移に関する一考察, 27th Fuzzy System Symposium, 2011.

[4] Haitham B. Ammar,Karl Tuyls, Matthew E. Taylor, Kurt Driessens, Gerhard Weiss, Reinforcement Learning Transfer via Sparse Coding, Proceedings of the 11th International Conferenceon Autonomous Agents and Multiagent Systems (AAMAS 2012), 4-8, 2012.

[5] Olshausen, B.A. and Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381:607-609, 1996.