2U-5

スパースコーディングを用いた強化学習における転移学習の一考察

齋藤　碧 †　　小林　一郎 †
† 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

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，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，新しいタスクで再利用することにより，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，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．

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 は割引率を表す．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．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．

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．

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \quad (3)$$

A Study on Transfer Learning with Sparse Coding in the framework for Reinforcement Learning

† Midori SAITO(saito.midori@is.ocha.ac.jp)
† Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
Advanced Sciences, Graduated School of Humanities and Sciences, Ochanomizu University (†)
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

ここで，右辺の第一項は $\boldsymbol{y}$ と復元された信号 $\boldsymbol{Dx}$ 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．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．

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 1 のように，スタート地点は (0, 14) であり，すべてのタスクで共通とした．また，ゴール地点を変えた 1〜4 と A，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)) について強化学習を行った．本実験では A と B を，1〜4 のどのタスクと似ているかの分類を行う．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 タスクにおいて得られたそれぞれの 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は 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た．この辞書を用いて，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．図 2 に 4 種類のタスクの係数ベクトルの結果を示す．また，図 3 は 1〜4 と同様にして取得した A と B の結果である．

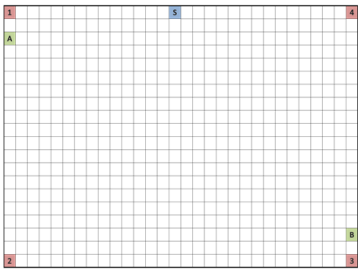

図 1: 20 × 30 マスの格子空間

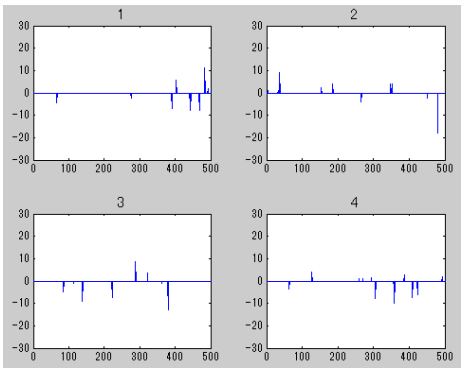

図 2: 係数ベクトル (1〜4)

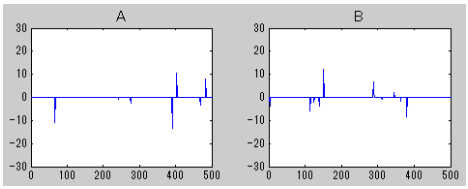

図 3: 係数ベクトル (A,B)

5.1 考察

表 1 に図 2 と図 3 の結果を平均絶対値誤差で表した．これより，タスク A はタスク 1 と，タスク B はタスク 3 と，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．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

	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，強化学習で得られた Q 値をスパースなベクトルに変形し，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による再現性を示した．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組み込みたいと考えている．

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.