6R-01

強化学習におけるスパースコーディングを用いた効率的な転移学習

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

転移学習において, 強化学習で得られた知識を転移し, 新しい環境にも適用することで, 最初から学習しなおすよりも学習を効率化することができる [1][2]. しかし, 強化学習の知識転移において, それらの蓄積された知識をどの程度の割合で転移するかを決定するのは難しい. そこで本研究では, スパースコーディング [3] を転移学習に適用することで知識の選択及び, その程度を明確にすることを可能にした新しい手法を提案する. 本実験では, コスト付き迷路における目的地までの最適経路を発見する課題において, 提案手法が従来の Q-learning と比較し, ターゲットタスクで良好な初期探索の実現と探索コストの削減に成功した.

2　強化学習の知識転移

2.1　強化学習

強化学習は，エージェントが環境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. 手順は以下の 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，選択可能な行動から一つ選び実行する. 3. その行動に対し，報酬または罰則を与えて評価する. 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，$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$ 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 $\pi(s, a) = Pr\{a_t = a | s_t = s\}$ によって表わされ，強化学習では報酬期待値を最大にする行動規則 $\pi^*(s, a)$ の獲得を目標とする.

2.2　Q-learning

本研究では強化学習の手法として Q-learning を採用した. Q-learning は，Q 値と呼ばれる状態と行動の評価値を最大化する. Q 値の更新式を以下に示す.

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

式 (1) 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，γ は割引率を表す.

2.3　転移学習

転移学習では，ソースタスクで強化学習により得られた方策や Q 値などの知識を，類似したターゲットタスクで事前知識として予め転移させておくことで，最初から学習しなおすよりも少ない探索回数で学習を行うことを目標とする. しかし，ターゲットタスクの状況によって, どの知識をどの程度転移させるかを正しく見極めないと負の転移が発生する可能性がある. そこで, 本研究では転移知識の選択にスパースコーディングを導入する.

3　スパースコーディング

転移知識を選択する際に, タスク間の類似度を考慮に入れる必要があり, 本研究では, その測定にスパースコーディング [3] を用いる. スパースコーディングは以下により定式化される.

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

式 (2) において $\boldsymbol{y}$ は入力信号を示しており，$\boldsymbol{D}$ 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.

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

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，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，最適なスパース係数行列 $\boldsymbol{x}^*$ が求められる.

4　スパースコーディングを用いた知識転移

この章では, 提案手法について説明する. 本研究では, 図 (1) にあるように, ソース・ターゲットタスク共に, 5 色のコスト付き (白:0, 青: -2, 緑:-3, 赤:-5, 黒:-10) の迷路 (縦:30, 横:30) をタスクとした. まず最初に, 複数のソースタスクで強化学習を行い, それぞれのタスクにて 900 マス分のマス目コストと Q 値を獲得した. 以下に提案手法の手順を示す.

Efficient Transfer Learning for Reinforcement Learning using Sparse Coding

† 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

step1 ターゲットタスクの現在探索しているマス目を含む周囲 25 マスのマス目コストを取得し, スパースコーディングの $\boldsymbol{y}$ に代入する

step2 スパースコーディングの $\boldsymbol{D}$ にも同様にソースタスクの 25 マス分のマス目コストを 1 基底として代入し, ターゲットの状態と類似しているソースの状態をスパースコーディングで算出する

step3 step2 の結果, $\boldsymbol{x}$ が非 0 の基底の添字に対応する部分の Q 値を $\boldsymbol{x}$ の割合で総和をとる

step4 step3 で求めた Q 値を, step1 で現在探索しているマス目へ転移する Q 値とする

step1〜step4 をターゲットタスク上の探索で繰り返すことにより (図 1), オンラインで Q 値を転移し, 学習効率の向上を図る.

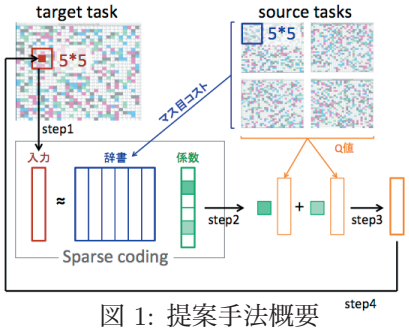

図 1: 提案手法概要

5 実験

本実験では, ソース・ターゲットタスク共に, 900 マスのコスト付き迷路を実験対象とし, スタートを左上, ゴールを右下に設定した. また, エージェントは上下左右に 1 マスずつ移動できるものとした. それぞれのタスクでは, マス目コストをランダムに再配置し, 新しい環境を用意した. このようなターゲット環境で, 100 エピソード繰り返し, それぞれのエピソードにかかったステップ数とコスト量を評価した.

5.1 実験 1：提案手法と Q-learning の比較

実験 1 では, 提案手法において, 4 つのソースタスク上で強化学習を行い, 4 つ分の辞書基底を作成した. 比較手法として, ターゲットタスクで転移知識なしの学習を Q-learning を用いて実行した. それぞれ 100 エピソードの 5 回分の平均をとった.

5.2 実験 1：実験結果および考察

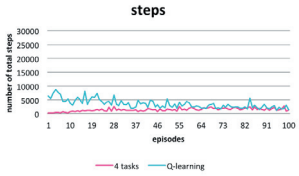

図 2: 実験 1:ステップ数

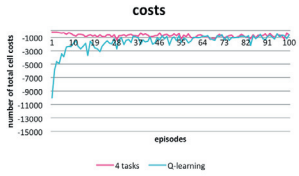

図 3: 実験 1:コスト量

図 2 と 3 に実験 1 の結果を示した. ここでは, 縦軸はそれぞれ 1 エピソードにかかったステップ数とマス目コスト, 横軸は 100 回分のエピソードを示している. また, 赤のグラフが提案手法, 青が Q-learning の結果である. これらより, 提案手法は, 学習しなおすよりもステップ数・マス目コストを少量に抑えてゴールに辿り着くことができた. しかし, 一部のターゲットマス目においては, 類似した基底が存在せず負の転移が起こったものもあった.

5.3 実験 2：提案手法における基底数の比較

実験 1 の考察をふまえ, 実験 2 では 10 ソースタスク分に辞書を拡張した. しかし, 10 タスク分の辞書では計算時間が増えるため, k-means 法を用いて 10 タスク分の辞書基底を 4 タスク分に基底数を制限した. 4 タスク分, 10 タスク分, k-means による基底数削減, の 3 つを同様に 5 回分の平均で比較した.

5.4 実験 2：実験結果および考察

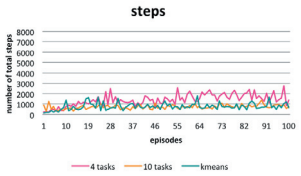

図 4: 実験 2:ステップ数

図 5: 実験 2:コスト量

図 4, 5 では, 赤が 4 タスク分, 橙が 10 タスク分, 緑が k-means の結果である. これより, 基底数の増加による転移精度の向上を確認した. また, k-means による基底数制限は, 4 タスク分より精度が低下した部分もあったが, これは k-means の結果で平均をとったため, Q 値をうまく転移できなかったと考えられる. 全体的には 10 タスク分と比べても転移精度を維持できた.

6 まとめ

本研究では, 強化学習の転移知識選択手法にスパースコーディングを導入した効果的な転移学習手法を提案した. 実験 1 より, 提案手法が知識選択を可能にし, Q-learning よりも転移精度向上に成功した. 実験 2 では, 提案手法の辞書基底数の変化による違いを検証した. そこで, k-means により基底数を制限しながらも転移精度を維持させることができた. 今後も更なる辞書精度向上を目指したいと考えている.

参考文献

[1] Trung Thanh Nguyen, Tomi Silander and Tze-Yun Leong, Transferring Expectations in Model-based Reinforcement Learning, NIPS, 2012.

[2] Haitham B. Ammar, Karl Tuyls, Matthew E. Taylor, Kurt Driessens, and Gerhard Weiss, Reinforcement Learning Transfer via Sparse Coding, AAMAS 2012, 4-8, 2012.

[3] 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.