3D3-OS-30a-2in2

逆強化学習を用いた転移可能な報酬関数の推定

Estimation of Transferable Reward Function via Inverse Reinforcement Learning

北里勇樹 *1　荒井幸代 *1
Kitazato Yuki　Arai Sachiyo

*1千葉大学大学院工学研究科都市環境システムコース
Graduate School of Engineering, Chiba University, Division of Urban Environment Systems

Though reinforcement learning has been popular for its applicability to control systems using minimum knowledge about their dynamics, it is impractical in real world applications because it takes long time to reach a certain policy from scratch. Transfer learning, in an RL setting, typically attempts to decrease training time by learning a source task before learning the target task. In this study, we propose a method to transfer reward function from the environment information using Inverse Reinforcement Learning. The effectiveness of the proposed method is verified by empirical experiments of maze problems.

1. はじめに

強化学習において学習効率の改善は重要な課題であり，そのための手法として，解決が容易な既知のタスクで得られた情報を未知のタスクに再利用する転移学習がある。これは過去に類似したタスク（以後，元タスク）を学習することで得た知識を，これから解こうとするタスク（以後，目標タスク）の学習に転用すべき知識を即座に獲得する，あるいはその知識を獲得するための手がかりを得ることによって，学習回数を削減する方法である。

転移学習という語は幅広い機械学習の枠組みとして用いられている。強化学習に応用した例として，fernardo ら [Fernando 06] は元タスクで学習した方策を，目標タスクを学習するエージェントに対して確率的なバイアスとして与え，学習エージェントが現在の方策を用いるか，新しい状態を探索するか，過去に学習した方策を用いるかを選択し，学習を進める手法を提案した。Matthew ら [Matthew 07] は元タスクと目標タスクが大きく異なる場合（例えば迷路問題から Keepaway 等）の転移を可能にするために，元タスクで得た方策をデータベースにまとめ，修正したものを目標タスクに用いるルールトランスファーという手法を提案した。これらの手法は元タスクで獲得した方策を転移させる手法である。

一方で方策ではなく報酬を転移させる手法として逆強化学習の分野でも研究されている。逆強化学習は最適な行動を知るエージェント（エキスパート）の行動軌跡と状態遷移確率を用いて報酬関数を推定する枠組みである。Monica ら [Monica 11] はそれぞれ別々の目的を持った複数のエキスパートの行動軌跡から同じ目的を持つエキスパートを EM アルゴリズムを用いてクラスタリングし，各クラスタに対して報酬関数を推定する。次に新たな軌跡に対してベイズルールからどのクラスタに含まれるかを推定し，報酬関数を割り当てる。Jaedeug ら [Jaedeug 12] はディリクレ過程混合モデルを利用し，あらかじめクラスタ数を設定する必要のないノンパラメトリックベイジアンアプローチによる報酬関数の推定を行った。この手法では新しい軌跡が与えられたときに，以前の学習結果を用いて効率的に報酬関数を計算する。逆強化学習を用いた転移学習は，所与として目標タスクにおけるエキスパートの行動軌跡が必要となるため，適用範囲が限定される。

そこで本研究では目標タスクにおいてエキスパートの行動軌跡の代わりに，環境の情報の一部を用いた報酬関数の転移法を提案する。具体的には元タスクとして，内部に複数の障害物が存在する迷路問題を用意し，それぞれの報酬関数を Abbeel らの逆強化学習 [Abbeel 04] により求める。次に各障害物の位置と報酬関数を入出力に設定した教師有学習を行うことにより，目標タスクに対する報酬関数を求める。最後に計算機実験により本手法の有用性の検証を行う。

連絡先: 北里勇樹，千葉大学大学院工学研究科都市環境システムコース，千葉市稲毛区弥生町 1-33，sutoratoo@gmail.com

1. ランダムに選んだ方策 π^0 から特徴期待値 $\mu^0 = \mu(\pi^0)$ を計算する
2. 各特長期待値に対する重み $w^1 = \mu_E - \mu^0$ を計算し，$i = 1$ とする
3. 終了条件 $t^i \leq \tau$ を満たすまで以下を繰り返す：
   (a) 報酬関数 $R = w^i \cdot \phi$ と強化学習を用いて，最適方策 π^i を求める
   (b) 最適方策 π^i から特徴期待値 $\mu^{(i)} = \mu(\pi^i)$ を計算し，$i = i + 1$ とする
   (c) エキスパートの特徴期待値への射影ベクトル $\bar{\mu}^{i-1}$ を計算する
   (d)$w^{(i)} = \mu_E - \bar{\mu}^{(i-1)}$，$t^{(i)} = ||\mu_E - \bar{\mu}^{(i-1)}||_2$ とする

図 1: Abbeel の逆強化学習のアルゴリズム

2. Abbeel の逆強化学習

各状態で最適な行動をとるエージェントをエキスパートと定義する。Abbeel の逆強化学習ではエキスパートの行動軌跡を所与とし，エキスパートと同じような行動軌跡が得られる報酬関数 R を推定する。具体的には，状態を $S \rightarrow [0,1]^k$ で表される特徴量 $\phi(s_t)$ で定義し，各特徴量の出現頻度によって計算される特徴期待値 $\mu(\pi) = E[\sum_{t=0}^{\infty} \gamma^t \phi(s_t)|\pi] \in R^k$ を用いて行動軌跡を数値化する。エキスパートの特徴期待値との差が ϵ 以下になるような報酬関数を推定することにより，エキスパートに近似した行動軌跡を獲得することができる。Abbeel の逆強化学習では，エキスパートの特徴期待値との差が τ 以下になる特徴期待値を学習を通して得ることができる報酬関数を推定する。なお，文献 [Abbeel 04] では max-margin 法と projection 法が提案されている。それぞれ QPsolver，正射影ベクトルの計算を用いて報酬関数を推定する。実験では projection 法が収束率の点でわずかに良い性能を示していたため，本研究では projection 法を用いる。projection 法のアルゴリズムを図 1 に示す。

パラメータ τ はエキスパートの特徴期待値と推定した報酬による特徴期待値の差が十分近づいたことを判定するものであり，終了判定を行うための閾値である。

なお，エキスパートの特徴期待値 μ_E は，エキスパートの m 試行の行動軌跡 $\{s_0^{(i)}, s_1^{(i)}, \ldots\}_{i=1}^m$ から式 (1) によって推定する。

$$\hat{\mu}_E = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{\infty} \gamma^t \phi(s_t^{(i)}) \tag{1}$$

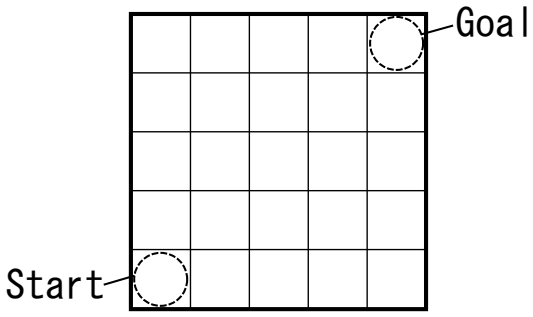

図 2: 実験環境

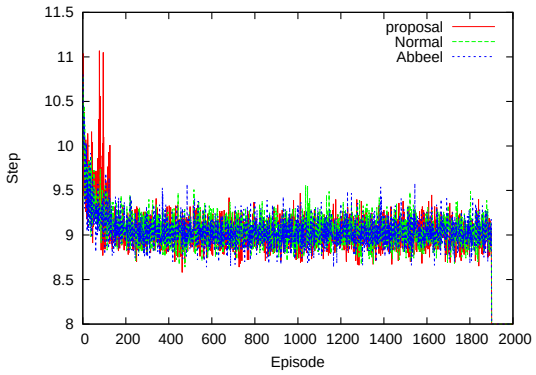

図 3: 学習曲線の比較

3. 提案手法

強化学習の学習効率を向上させるために，転移学習と逆強化学習を組み合わせ，環境の一部の情報から報酬関数を推定する手法を提案する。提案手法は，ステップ 1 : 元タスクにおける強化学習，ステップ 2 : 逆強化学習による報酬関数の推定，ステップ 3 : ニューラルネットワークによる目標タスクへの報酬の転移の三つのステップで構成される。各ステップについて迷路問題を例に挙げ，詳細に述べる。

まず，ステップ 1 では元タスクを用意する。ここではスタート，ゴール，障害物の座標を所与として，ゴールに対して報酬を与え，各迷路問題に対する最適方策を求め，最適方策からエキスパートの行動軌跡を求める。次にステップ 2 では，ステップ 1 で求めたエキスパートの行動軌跡を所与として，Abbeel の逆強化学習を行い，各状態に関する詳細な報酬関数を推定する。ステップ 3 では，元タスクの環境の情報を入力，ステップ 2 で求めた報酬関数を出力としたニューラルネットワークを作成し，バックプロパゲーションを用いてニューラルネットワークの各重みを計算する。

以上の 3 ステップを通して作成したニューラルネットワークに対して，目標タスクの環境の情報を入力することによって，目標タスクにおける報酬関数を推定する。

4. 計算機実験

図 2 に示す環境を用いて，提案手法の評価を行う．図 2 は，スタートを座標 (0,0)，ゴールを座標 (4,4) に持つ迷路問題である．今回の実験設定を次に説明する。まず元ドメインとなる迷路を生成する。これは図 2 を基本として，スタートとゴールの位置は座標 (0,0)，ゴールを座標 (4,4) とする。次に障害物を生成する。この障害物は 3 から 10 個とし，スタートとゴール以外の座標に対してランダムに設定した。なおゴールにたどり着けない場合は除外して，10 個の迷路問題を作成した。次に各迷路問題に対して正解の行動軌跡をそれぞれ強化学習を用いて獲得した。次に生成した迷路問題の報酬関数を逆強化学習を用いて求める。逆強化学習のパラメータは割引率を 0.9，学習率を 0.03 とし，繰り返し回数の上限を 100 に設定した。各試行で得られた報酬によるエージェントの軌跡が，エキスパートと一致した時点で終了とした。次にニューラルネットワークでは学習率を 0.1，初期重みを [-0.1,0.1] の範囲でランダムに作成し，バックプロパゲーションによる学習の繰り返し回数を 1000 とした。

実験の評価には，(a) 最短経路が右回り，(b) 最短経路が左回り，(c) 最短経路が中央を通るものの三つを用いた。実験の結果，(a) の迷路問題のみ，転移した報酬関数で学習できた。100 試行の学習曲線の平均値を図 3 に示す。ここで Normal はゴールに報酬値 0，その他の状態に報酬値-1 を設定したものであり，Abbeel は Abbeel の逆強化学習で報酬を設定したものである。

図 3 より，3 つの学習曲線に有意な差を見て取れない。

次に結果の考察を行う。まず，今回提案した手法で，迷路問題の (b)，(c) で最短経路を獲得可能な報酬関数を推定できなかった理由としては，訓練データ数が少なかったことや各パラメータの調整不足が考えられる。これは今後追加実験を行い検証していく予定である。次に各手法の学習曲線に差が見られなかった理由について考察する。これは Abbeel の逆強化学習が Normal の学習曲線とほとんど同じであることから，Abbeel の逆強化学習は学習効率を改善するための報酬関数を推定するわけではなく，エキスパートと同一の行動軌跡を獲得するためのものであることがわかる。したがって，今後は転移させる報酬関数は，学習効率を改善する報酬関数を選ぶ必要があることがわかる。

5. まとめ

強化学習の学習効率を向上させるために，転移学習に逆強化学習と教師有学習を組み合わせた手法を提案した。これは，通常の転移学習では元タスクから方策を転移させるが，今回提案した手法では報酬関数を転移させるためのものである。今回の実験では，各パラメータや訓練データ等の調整が不十分であったため，期待する結果が得られなかった。今後の方針としては，各パラメータやデータ数を増やし，今回提案した手法の性能を正確に評価することと，転移させる報酬関数に学習効率の改善を陽に取り入れることである。

参考文献

[Monica 11] Monica Babes-Vroman , Vukosi Marivate, Kaushik Subramanian, Michael Littman : Apprenticeship Learning About Multiple Intentions, In Proc. ICML, 2011.

[Jaedeug 12] Jaedeug Choi and Kee-Eung Kim : Nonparametric Bayesian Inverse Reinforcement Learning for Multiple Reward Functions, In Proc. NIPS, pp.1-9, 2012.

[Abbeel 04] P. Abbeel, and A. Ng: Apprenticeship learning via inverse reinforcement learning, Proceedings of the 21th International Conference on Machine Learning, ICML ’ 04, 2004.

[Fernando 06] Fernando Fernandez, Manuela Veloso : Probabilistic Policy Reuse in a Reinforcement Learning Agent, Proceedings of the fifth international joint conference on Autonomous Agents and Multiagent Systems, pp.720-727,2006.

[Matthew 07] Matthew E. Taylor and Peter Stone : Cross-domain transfer for reinforcement learning. In Proceedings of the Twenty-Fourth International Conference on Machine Learning, June 2007.