逆強化学習における非演示軌道を利用した特徴期待値の算出

○五味 達朗（中央大学)， 増山 岳人（中央大学)， 梅田 和昇（中央大学）

Feature Expectation in Inverse Reinforcement Learning Using Non-Demonstration Trajectories

○Tatsurou GOMI (Chuo Univ.), Gakuto MASUYAMA (Chuo Univ.) , and Kazunori UMEDA (Chuo Univ.)

Abstract: This paper presents model-free inverse reinforcement learning technique using non-demonstration trajectories. The non-demonstration trajectories are supposed to be sampled from arbitrary policy of a learner. We utilize the sampled trajectories for estimation of reward function in semi-supervised manner to reduce the number of required demonstrations. Simulation results in highway problem indicated that proposed method could improve the performance of inverse reinforcement learning.

１．緒言

未知環境下におけるロボットの行動学習手法として，強化学習についての研究が盛んに行われてきた．強化学習では，試行錯誤的な経験の収集から将来に渡って得られる報酬の期待値を最大化する方策を獲得する．その際，学習者であるエージェントは環境から与えられる報酬と呼ばれる信号を頼りに学習を行う．しかし，複雑なタスク，環境においては報酬関数の設計が困難であるという問題点が挙げられる．

そこで，その問題点を解決するために，報酬関数をエキスパートの演示から推定する逆強化学習に関する研究が行われてきた．逆強化学習により推定された報酬関数を用いて強化学習を行うことで，エキスパートの振る舞いを再現する模倣学習が可能となる [1]．

Auddifren らは特徴期待値のマッチングに基づく逆強化学習手法と非演示軌道を利用した，半教師ありの逆強化学習を提案している [2]．この手法では，非演示軌道をエキスパートの特徴期待値の算出に利用することで学習結果を向上させることに成功している．しかし，Auddifren らの手法では，状態遷移確率等の環境モデルが要求される．

そこで，本稿では Auddifren らと同様のアプローチにより，環境モデルを必要としない半教師あり逆強化学習手法を提案する．シミュレーション実験により，非演示軌道を用いることで，よりよい方策が学習可能であることを確認する．

２．非演示軌道を利用した特徴期待値の算出

本稿では，逆強化学習にはモデルフリーである Relative Entropy Inverse Reinforcement Learning（REIRL）[3]を用いる．まず，エキスパート及びエージェントの取りうる状態をs，とりうる行動をaとする．方策$\pi(s,a)$は状態sにおいて行動aが選択される確率である．状態sにおいて行動aを選択し観測される特徴量を$\boldsymbol{f}(s,a)$とすると，初期状態s_0から方策πに従って行動した場合の特徴期待値は$\boldsymbol{\mu}^{\pi} = E[\sum_{t=0}^{\infty}\gamma^t \boldsymbol{f}(s_t,a_t)|s_0,\pi]$となる．また，パラメータベクトルを$\boldsymbol{\theta}$とする．ここで，報酬関数を$\boldsymbol{\theta}$と$\boldsymbol{f}$の線形結合$r(s,a) \coloneqq \sum_{i=1}^{k}\theta_i f_i(s,a)$とすると，状態$s$において方策$\pi$に従ったときの価値関数は$V^{\pi}(s) = \sum_{i=1}^{k}\theta_i \mu_i^{\pi}(s,a)$となる．

$\|\boldsymbol{\theta}\| = 1$とした場合，エキスパートとエージェント間の価値関数の相違度は特徴期待値を用いて表すことができる．REIRL では，パラメータベクトル$\boldsymbol{\theta}$の値を，

$$|\sum_{\tau\in T} p_{\tau}(\tau)\mu_i^{\tau} - \hat{\mu}_i| \le \epsilon_i \tag{1}$$

の拘束のもと最適化する．ここで，$\hat{\mu}_i$はエキスパートの演示軌道から得られた特徴期待値である．また，$p_{\tau}(\tau)$は軌道τがとられる確率であり，μ_i^{τ}は軌道τに沿って観測される特徴量の期待値である．演示軌道の集合を$\Lambda^E = \{\tau_i^E\}_{i=1}^{l}$，非演示軌道の集合を$\Lambda = \{\tau_j\}_{j=1}^{u}$とする．また，それらすべての軌道の集合を$\Lambda^* = \{\tau_k^*\}_{k=1}^{l+u}$とする．

本稿では，特徴期待値を以下のように算出する．

$$\mu^* = \frac{1}{l(l+u)}\frac{\sum_{l=1}^{l}\sum_{k=1}^{l+u}\phi(\tau_l^E,\tau_k^*)\boldsymbol{f}(\tau^*)}{\sum_{l=1}^{l}\sum_{k=1}^{l+u}\phi(\tau_l^E,\tau_k^*)} \tag{2}$$

ここで，ϕは軌道間の類似度を算出する関数であり，以下のように定義される．

$$\phi(\tau,\tau') = \exp(-\|f(\tau) - f(\tau')\|^2/2\sigma^2) \tag{3}$$

式(2)の特徴期待値は，エキスパートの演示軌道だけでなく，式(3)で与えられる類似度によって重み付けられた非演示軌道を半教師あり学習的に用いることで算

出される．そのため，類似度が正しくそれぞれの非演示軌道の特徴期待値算出への寄与を表していれば，推定される報酬関数及び方策は，よりエキスパートの振る舞いを忠実に再現するものとなることが期待できる．

3．シミュレーション実験

提案手法の有効性を検証するため，逆強化学習手法の評価に広く用いられる Highway problem[1]によるシミュレーションを行った．1）演示軌道のみを用いて特徴期待値を算出する REIRL; 2）演示軌道に加えて非演示軌道を特徴期待値の算出に利用する提案手法; 上記 2 手法について獲得される方策の性能を評価した．また，方策の算出には価値反復法を用いた．

3.1　実験設定

Highway problem では，Fig.1 のような環境で自動車を走行させる．状態は自車の座標x_m，y_m及びy軸方向の速度v_m，他車の座標x_o，y_o及びy軸方向の速度v_oの組み合わせによって定義した．ただし，y_mおよびv_oは一定とし，第 3，5，7 レーンのいずれかに常に 1 台のみ他車が存在し，直進しているものとした．v_oの速度で 1 ステップあたりy軸方向に 1 進むとし，自車の速度は$2v_o$，$3v_o$，$4v_o$の 3 段階とした．また，行動は何もしない，左右へ 1 レーンの移動，速度の加減速の 5 つとした．

真の報酬は自車が側道（第 1，2，8，9 レーン）を走っている場合 - 0.5，他車が自車の周囲 9 マス内に存在している場合 - 1，自車が車道（第 3 から 7 レーン）を走っており，かつ最高速度を出している場合 1 を得るように設定した．

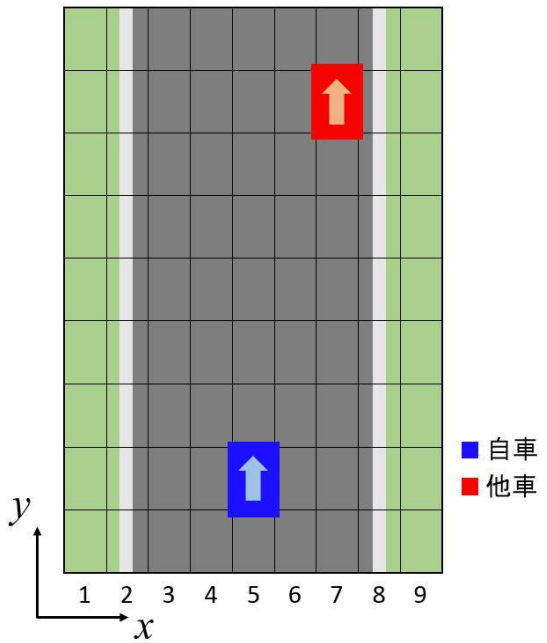

Fig.1　Highway problem の環境

Table 1　実験結果

	REIRL	提案手法
累積報酬	50.11	76.68

エキスパートの演示は真の報酬から価値反復法によって得られた最適方策からサンプルした．演示軌道数は 5，非演示軌道数は 100，σは 0.15 とした．

3.2　実験結果

実験結果の評価には 100 ステップ走行した際の累積報酬の平均値を用いた．100 試行のシミュレーションを行った結果の平均を Table 1 に示す．

実験結果より，提案手法は比較手法の REIRL より約 26.5 多くの報酬を得ることができている．演示軌道に類似した非演示軌道を利用することで，モデルフリー逆強化学習の性能向上が可能であることが示唆された．

4．結言

本稿では，特徴期待値の算出に非演示軌道を用いたモデルフリーの逆強化学習を提案し，シミュレーション実験により手法の有用性を評価した．

今後の展望として，より複雑な制御対象への適用が挙げられる．本稿では状態数の比較的少ない問題に対して本手法を適用している．しかし，状態数が少ないということは類似度が近くなる確率が高いことを示しており，今後は状態数の多い学習対象に対してどのように非演示軌道を収集するかが課題となる．

参考文献

1) P. Abbeel, A. Ng: “Apprenticeship learning via inverse reinforcement learning,” Proc. of the 21st ACM Int. Conf. on Machine Learning, 2004.

2) J. Audiffren, M. Valko, A. Lazaric, M. Ghavamzadeh: “Maximum entropy semi-supervised inverse reinforcement learning,” Proc. of the 23rd Int. Joint Conf. on Artificial Intelligence, 2015.

3) A. Boularias, J. Kober, J. Peters: “Relative entropy inverse reinforcement learning,” Proc. of the 14th Int. Conf. on Artificial Intelligence and Statistics, pp.182-189, 2011.