H-022

軌道のスコアに基づく逆強化学習を用いた視覚演示からの報酬関数の推定

Estimation of Reward Function from Visual Demonstrations Using Inverse Reinforcement Learning from Scored Trajectories

江尻 尚馬[1)]　福永 修一[2)]　黒木 啓之[1)]
Shoma Ejiri　Shuichi Fukunaga　Takashi Kuroki

1　はじめに

逆強化学習は，エキスパートの行動データから，目的のタスクをどれだけ達成したかを示す指標となる報酬関数を推定する機械学習の一つの方法である [1]. しかし，コストの問題で行動データが直接観測できない場合がある．そのため，視覚演示のみを用いた逆強化学習として，Cross- embodiment Inverse Reinforcement Learning (XIRL) [2] という方法が提案されている．XIRL はニューラルネットワークによって学習した特徴量を使うことで異なる角度からの撮影や異なる環境での実施形態においても一貫した報酬関数を推定する．具体的には，人間によるタスクの視覚演示から，マニピュレータの強化学習に用いることのできる報酬関数を推定を行うことができた.

しかし，この方法では事前学習に最適，または準最適な軌道を必要とし，最適でない軌道を含んでいた場合に推定精度が悪くなることがシミュレーション結果から示されている．一方で，軌道のスコアに基づく逆強化学習 [3, 4] という方法が提案されている．この方法では最適，または準最適な軌道ではない軌道を含む任意の軌道とその軌道に対して評価者によるスコアのペアから報酬関数を推定する．この方法では最適でない軌道がデータに含まれていても報酬関数を推定することができた.

本研究は軌道のスコアに基づく逆強化学習を用いた視覚演示からの報酬関数の推定を目的とする．状態表現が抽象的な視覚演示から畳み込みニューラルネットワークによる特徴抽出を行い，軌道のスコアに基づく逆強化学習を用いることで，最適でない軌道を含む任意の軌道から報酬関数を推定する．提案手法を平面 2 リンクマニピュレータの物理シミュレーションに適用し，最適でない軌道が含まれたデータから学習できることを示す.

2　ResNet による特徴量の学習

Residual Network（ResNet）[5] は He らによって提案された畳み込みニューラルネットワークのモデルである．通常は物体検知や分類タスクに使われるが，本研究では特徴抽出器として扱う．画像 $s \in S$ に対する ResNet から得られる出力を特徴量 $\boldsymbol{\phi} : S \mapsto \mathbb{R}^k$ とする．この時，S は画像全体の集合である．画像 s を入力とした特徴マップによる非線形変換からベクトル $\boldsymbol{g} \in \mathbb{R}^C$ を得られる．ここで，C はチャネル数である．ベクトル $\boldsymbol{g}$ を用い m 次元の特徴量 $\boldsymbol{\phi} \in \mathbb{R}^k$ への線形変換を行う.

$$\boldsymbol{\phi}(s) = \boldsymbol{W}\boldsymbol{g}(s) + \boldsymbol{b} \tag{1}$$

ここで $\boldsymbol{W} \in \mathbb{R}^{k \times C}$ は重み行列であり，$\boldsymbol{b} \in \mathbb{R}^k$ はバイアスベクトルである．重み行列とバイアスベクトルは ResNet の学習によって求めることができる.

1）　東京都立産業技術高等専門学校
Tokyo Metropolitan College of Industrial Technology
2）　広島工業大学
Hiroshima Institute of Technology

3　提案手法

本研究では，ResNet から得られる特徴量を用いて，軌道のスコアに基づく逆強化学習を適用し，軌道と与えられるスコアから報酬関数を得る手法を提案する．ここでの軌道は任意のタスクを行った動画であり，スコアは評価者が与える目的のタスクにどれだけその軌道が適しているかを評価する値である．報酬関数は特徴量の線形結合モデルで表すことができるため，報酬 R を重み $\boldsymbol{\psi} \in \mathbb{R}^k$ と 2 節で扱った ResNet から得られる特徴量 $\boldsymbol{\phi}$ の積として以下のように表す.

$$R(s) = \boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\phi}(s) \tag{2}$$

終端時刻 T までの時系列順の状態遷移を持つ i 本目の軌道を $\tau_i = \{s_0^i, ..., s_T^i\}$ とする．軌道 τ_i で得られるスコア v_i をその軌道の報酬の累積和として

$$v_i = \sum_{t=0}^{T} \gamma^t R(s_t^i) \tag{3}$$

と表す．ただし，$\gamma \in [0, 1]$ は時系列毎の割引率である．式 (3) に式 (2) を代入すると，v_i は以下の式で表せる.

$$v_i = \boldsymbol{\psi}^{\mathrm{T}} \sum_{t=0}^{T} \gamma^t \boldsymbol{\phi}(s_t^i) \tag{4}$$

l 本の軌道とそれらに付与された真のスコア v^* がそれぞれ $(\tau_1, v_1^*), ..., (\tau_l, v_l^*)$ と与えられたとき，以下の目的関数 $J(\boldsymbol{\psi})$ を考える.

$$J(\boldsymbol{\psi}) = ||\boldsymbol{M}\boldsymbol{\psi} - \boldsymbol{v}^*||^2 \tag{5}$$

ここで，$\boldsymbol{v}^* = (v_1^*, ..., v_l^*)^{\mathrm{T}} \in \mathbb{R}^l$ は軌道のスコアを要素とするベクトルであり，$\boldsymbol{M} \in \mathbb{R}^{l \times k}$ は以下の式で定義さ

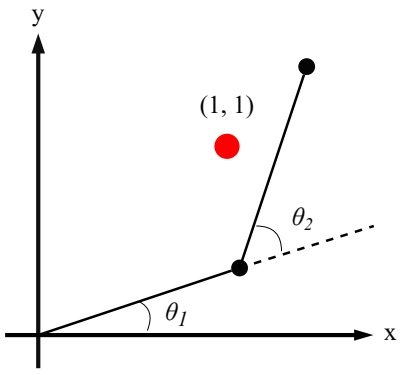
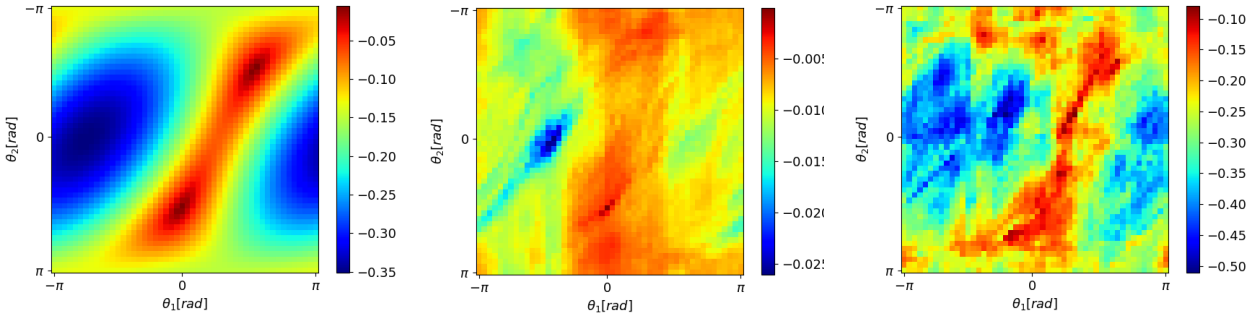


図 1: 平面 2 リンクマニピュレータ



(a) 真の報酬関数　(b) XIRL　(c) 提案手法

図 2: 報酬関数の推定結果

れる．

$$\boldsymbol{M} = \begin{pmatrix} \sum_{t=0}^{T} \gamma^t \boldsymbol{\phi}(s_t^1)^{\mathrm{T}} \\ \vdots \\ \sum_{t=0}^{T} \gamma^t \boldsymbol{\phi}(s_t^l)^{\mathrm{T}} \end{pmatrix} \tag{6}$$

このことより，式 (5) を最小にする最小二乗解 $\hat{\boldsymbol{\psi}}$ は以下の式で求められる．

$$\hat{\boldsymbol{\psi}} = (\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})^{-1}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{v}^* \tag{7}$$

4　数値例

本節では，既存手法である XIRL と提案手法をシミュレーション上で比較することで，提案手法の有用性を確認する．平面 2 リンクマニピュレータでのタスクの目的は，図 1 に示すようにランダムな初期位置から平面 2 リンクマニピュレータの手先を目標点にできるだけ近づけることである．θ_1, θ_2 は平面 2 リンクマニピュレータのそれぞれの関節角であり，目標点の座標は $(1, 1)$ とした．シミュレーション環境には，物理シミュレータ MuJoCo の Reacher[6] を使用した．

ここでの真の報酬関数は，マニピュレータの手先から目標点までの負の距離とした．特徴量として ImageNet の画像認識タスクの学習済みモデルである ResNet18 の中間層から得られる 32 次元の特徴量を使用した．また，ステップ数である終端時刻 $T = 10$ で軌道数 $l = 1000$ のランダムな軌道を用いて推定を行った．提案手法に与えるスコアはランダムな軌道の真の報酬関数から与えられる 1 エピソードあたりの報酬の累積和とした．軌道に対するスコア $v_1^*, ..., v_l^*$ は以下の式で定義する．

$$v_i^* = \sum_{t=0}^{T} \gamma^t \left(-\sqrt{{d_x^{(t)}}^2 + {d_y^{(t)}}^2} \right) \tag{8}$$

ここで，t はステップ数であり，d_x, d_y はそれぞれマニピュレータの手先から目標点までの距離である．

図 2(a), 2(b), 2(c) に真の報酬関数と XIRL によって推定した報酬関数，軌道のスコアに基づく逆強化学習によって推定した報酬関数を示す．図 2(a), 2(b), 2(c) より，提案手法と XIRL から推定した報酬関数を比較すると，提案手法から推定した報酬関数がより真の報酬関数に近いことが確認できる．また，それぞれの推定した報酬関数と真の報酬関数との二乗平均平方根誤差 (RMSE) を求めた．XIRL の RMSE は 0.196 となり，提案手法の RMSE は 0.107 となった．RMSE の比較からも，提案手法がより正確であることが示された．これらの結果から，提案手法が精度よく報酬関数を推定できたことを確認した．

5　おわりに

畳み込みニューラルネットワークの学習済みモデルを用いて軌道のスコアに基づく逆強化学習による視覚演示から報酬関数の推定する方法を提案した．状態表現が抽象的な視覚演示から畳み込みニューラルネットワークによる特徴抽出を行い，軌道のスコアに基づく逆強化学習を用いることで，最適でない軌道を含む任意の軌道から報酬関数を推定した．提案手法を Reacher に適用し，最適でない軌道が含まれたデータから学習できることを示した．

今後の研究の課題として，推定した報酬関数を用いて強化学習を行い，目的のタスクを達成する行動を得られるかを確認することが挙げられる．また，XIRL では Temporal Cycle-Consistency(TCC)[7] を利用して，異なる実施形態においても報酬関数を推定していた．軌道のスコアに基づく逆強化学習においても TCC エンコーダを学習することで，異なる実施形態でも学習が行えるように改善することができると考えられる．

参考文献

[1] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. Artificial Intelligence, Vol. 297, p. 103500, 2021.

[2] Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. XIRL: Cross-embodiment inverse reinforcement learning, 2021.

[3] Layla El Asri, Romain Laroche, and Olivier Pietquin. Reward shaping for statistical optimisation of dialogue management, Statistical Language and Speech Processing, pp. 93–101, 2013.

[4] Benjamin Burchfiel, Carlo Tomasi, and Ronald Parr. Distance minimization for reward learning from scored trajectories. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, No. 1, Mar. 2016.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[6] Mark Towers et al., Gymnasium, 2023. doi: 10.5281/zenodo.11232524.

[7] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June, 2019.