順序尺度型の一対比較データを用いた逆強化学習

Inverse reinforcement learning
using ordinal paired comparison data

数学専攻　中央大学大学院 理工学研究科 数学専攻
酒折研究室 修士 2 年 坂本 佳紀

1　はじめに

近年, デジタル化の進展や計算機の性能向上により多様で膨大なデータを得ることが可能になり, 行動選択ルールの最適化を扱う強化学習は多くの領域で活用されるようになった. 強化学習は深層学習と組み合わせることにより, 囲碁やビデオゲームにおいてはプロを超えるパフォーマンスを達成した. その他にも医療統計分野における治療 (介入) を行動として強化学習の枠組みを用いる動的治療計画の研究やビジネスにおける交渉モデルなど, 多岐にわたる領域での応用が期待されている.

強化学習では事前に報酬関数を設定し, 与えられた環境の下で報酬を最大とするような行動をとるようにエージェント (学習者) を学習させていく. しかし, 現実の複雑なタスクでは, 状態空間が大きいだけでなく, 目的に関係する要因を報酬関数として表現することが困難な場合がある. 報酬関数の設計が不適切だと, 我々が意図しない行動を学習してしまったり, 学習の効率が著しく低下する恐れがある.

逆強化学習は, 報酬を手動で設計せず, そのタスクにおいて最適な行動を実行できるような熟練者の存在を仮定し, 熟練者が生成した状態行動対の軌跡を教師データとして報酬を推論するアルゴリズムである. 逆強化学習においても深層学習と組み合わせることにより, ロボットの歩行やシミュレーション環境での運転など高次元のタスクでも学習が可能となった. しかし, 逆強化学習による報酬関数の学習には最適な行動をすることで得られる軌跡が必要であったため, 熟練者の行動を収集することが難しいタスクでは学習できなかった. この問題に対し, さまざまな質の軌跡から作成した一対比較データを用いて報酬を推論する T-REX [2] などのモデルが提案され, 必ずしも熟練していないエージェントによる教師データからも学習を進めることができるようになった.

T-REX の学習に使用する一対比較データのラベルは, 二つの軌跡の優劣しかつけられなかった. そこで, ラベルを順序尺度に拡張し, 引き分けなど多段階のラベル付けに対応した逆強化学習モデルを提案する.

2　強化学習

強化学習とは, ある環境内におけるエージェント (学習者) が, 現在の状態を観測し, 取るべき行動を決定する問題を扱う機械学習の一種である. 強化学習は学習者自身が環境を探索することで主体的にデータを獲得し, 得られた環境の情報をもとに最適な行動を決定することが目的である.

強化学習では行動選択ルールの最適化を扱うため, 状態の確率変数 S のみの確率過程 $\{S_t | t \in \mathbb{N}_0\} = S_0, S_1, \dots$ ではなく, 行動の確率変数 A などを追加した確率制御過程 $\{S_t, A_t | t \in \mathbb{N}_0\} = S_0, A_0, S_1, A_1, \dots$ を考える. マルコフ連鎖に行動と報酬を組み入れた確率制御過程をマルコフ決定過程 (Markov decision process, MDP) と呼び, 以下の 5 つ組 $M = \{\mathcal{S}, \mathcal{A}, p_{s_0}, p_T, R\}$ で構成される.

有限状態集合　$\mathcal{S} = \{s_1, \dots, s_{|\mathcal{S}|}\}$
有限行動集合　$\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$
初期状態確率関数　$p_{s_0} : \mathcal{S} \to [0, 1], \quad p_{s_0} = \Pr(S_0 = s)$
状態遷移確率関数　$p_T : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1], \quad p_T(s' | s, a) = \Pr(S_{t+1} = s' | S_t = s, A_t = a), \quad t \in \mathbb{N}_0$
報酬関数　$R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

状態遷移確率関数 p_T は現在の状態 s と行動 a のみに依存する．報酬関数は有界であり, 状態のみに依存する報酬 $R : \mathcal{S} \to \mathbb{R}$ や, 次状態にも依存する報酬 $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ などを用いる．強化学習では多くの場合, 環境をマルコフ決定過程でモデル化する．

行動の選択ルールを規定する関数である方策は, 現時間ステップの状態 s のみに依存して確率的に行動を選択する確率的方策 $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$

$$\pi(a|s) = \Pr(A = a|S = s)$$

を主に用いる．ある方策 $\pi(\cdot|s)$ を環境に入力することにより, データ $\{s_0, a_0, r_0, s_1, a_1, r_1, \ldots\}$ が収集される．

強化学習の目的関数は, 一般的には以下の条件付き期待リターン

$$V^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s\right]$$

を用いる．ここで $\gamma \in [0, 1)$ は割引率と呼ばれ, 長期的な報酬和をどの程度考慮するかを調整するパラメータである．割引累積報酬を考えることにより, エージェントが直近で得られる報酬を優先するようになる．任意の初期状態 $s \in \mathcal{S}$ からの条件付き期待リターンを最大化する方策

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, V^{\pi}(s)$$

を最適方策という．強化学習では, エージェントが環境を探索することで得た報酬や次状態のデータを用いて各状態の条件付き期待リターンを推定し, それを用いて方策 π の評価や最適方策 π^* の計算をする．

3 逆強化学習

逆強化学習は, 報酬の設計が困難な問題における報酬推論アルゴリズムである．強化学習では与えられた (自分で設計した) 報酬関数 R をもとに最適な方策 (行動) π を推定するが, 逆強化学習では報酬を定義する代わりに最適方策を実行できる熟練者の存在を仮定し, 熟練者の状態行動 (最適方策) の軌跡 $\tau = \{(s_0, a_0), (s_1, a_1), \ldots\}$ から, 報酬関数 $R(s)$,$R(s, a)$ を推定する．

3.1 Trajectory-ranked Reward EXtrapolation (T-REX)

一般的な逆強化学習の目的は, 熟練者による教師データの行動が最適となるような報酬関数の推論である．しかし, タスクによっては最適な行動を集めるのは難しい場合がある．Trajectory-ranked Reward EXtrapolation (T-REX) [2] では教師データの最適性を仮定しない代わりに, 複数のエージェントが生成した軌跡や, あるエージェントが時間をかけてタスクを練習する過程で生成した軌跡などを使用する．収集された軌跡 $\tau_i, (i = 1, \ldots, m)$ から軌跡のペア (τ_i, τ_j) をランダムに取り出し, 何らかの基準で比較し優劣を与え, 一対比較データを作成する．T-REX は軌跡の一対比較データを用いて報酬を推論する．

T-REX は軌跡の優劣を説明できるような報酬関数, つまり優れている軌跡に大きい報酬を与える報酬関数を得ることが目標である．軌跡の優劣を説明できるような報酬関数を求めれば, 学習した報酬関数のもとで強化学習を実行することで教師データより大きな真の報酬が得られる方策を学習できる可能性がある．具体的には報酬関数の予測器 $r_\theta : \mathcal{S} \to \mathbb{R}$ にニューラルネットワークを用いて, $i < j$ のとき $\sum_{s \in \tau_i} r_\theta(s) < \sum_{s \in \tau_j} r_\theta(s)$ となるように損失

関数 $\mathcal{L}(\theta)$ を

$$\begin{aligned}\mathcal{L}(\theta) &= -\sum_{\tau_i \prec \tau_j} \log \Pr\left(\sum_{s\in\tau_i} r_\theta(s) < \sum_{s\in\tau_j} r_\theta(s)\right) \\ &\approx -\sum_{\tau_i \prec \tau_j} \log \frac{\exp\left\{\sum_{s\in\tau_j} r_\theta(s)\right\}}{\exp\left\{\sum_{s\in\tau_i} r_\theta(s)\right\} + \exp\left\{\sum_{s\in\tau_j} r_\theta(s)\right\}}\end{aligned}$$

と設定する．この損失関数は一対比較データの分析手法である Bradley-Terry モデルに基づいている．

3.2 提案モデル

T-REX は二つの軌跡 τ_i, τ_j のラベル付けの際, 必ず優劣をつけなければならなかった．しかし二つの軌跡の真のリターンがほとんど同じであり, 優劣がない, もしくはわからないと判断したい場合がある．逆に軌跡のリターンに大きな差があると判断できる場合もある．そこで, 順序尺度を用いた Bradley-Terry モデル [1] に基づいて T-REX の損失関数を変更することで一対比較データの多段階のラベル付けに対応した逆強化学習モデルを提案する．

順序尺度のカテゴリ数を C とする．軌跡 τ_i, τ_j に対し, $\Pr(Y_{ij} = c),\ c = 1, \ldots, C$ は比較の結果がカテゴリ c になる確率を表す．c が小さいほど τ_j にとって有利な結果であり, c が大きいほど τ_i にとって有利な結果である．例えば $C = 5$ とすると, $c = 1$ のとき τ_j が大きく有利, $c = 2$ のとき τ_j が少し有利, $c = 3$ のとき引き分け, $c = 4$ のとき τ_i が少し有利, $c = 5$ のとき τ_i が大きく有利を表す．

二つの軌跡 τ_i, τ_j に対し, 潜在的な連続型確率変数 Y^*_{ij}, Y_i, Y_j が存在し, $Y_{ij} = c \Leftrightarrow \alpha_{c-1} < Y^*_{ij} < \alpha_c, Y^*_{ij} = Y_i - Y_j$ と表せると仮定する．ここで, Y_i, Y_j はそれぞれパラメータ μ_i, μ_j をもつ分布に従う確率変数である．μ_i は軌跡 τ_i の良さを表すパラメータであり, τ_i の累積報酬 $\mu_i = \sum_{s\in\tau_i} r_\theta(s)$ とする．$\alpha_0, \ldots, \alpha_C$ はカットポイントと呼ばれる実数を C 個の区間に分割するパラメータであり, τ_i, τ_j の良さの差を表す確率変数 Y^*_{ij} が $\alpha_{c-1} < Y^*_{ij} < \alpha_c$ であるとき, 比較の結果がカテゴリ c であることを意味する．

このとき, 比較の結果がカテゴリ c 以下である累積確率 $\Pr(Y_{ij} \le c)$ は Y^*_{ij} の累積分布関数 F を用いて

$$F^{-1}(\Pr(Y_{ij} \le c)) = \alpha_c - \left(\sum_{s\in\tau_i} r_\theta(s) - \sum_{s\in\tau_j} r_\theta(s)\right)$$

と表せる．F^{-1} にロジットリンクを適用することで累積ロジット Bradley-Terry モデル

$$\log \frac{\Pr(Y_{ij} \le c)}{\Pr(Y_{ij} > c)} = \alpha_c - \left(\sum_{s\in\tau_i} r_\theta(s) - \sum_{s\in\tau_j} r_\theta(s)\right)$$

が得られる．

軌跡ペア (τ_i, τ_j) と順序尺度のラベル $y_{ij} = c$ が与えられたとき, $r_\theta(s)$ のパラメータ θ とカットポイント

$\alpha_1, \ldots, \alpha_{C-1}$ の損失関数 $\mathcal{L}(\theta, \alpha_1, \ldots, \alpha_{C-1})$ を負の対数尤度

$$
\begin{aligned}
&\mathcal{L}(\theta, \alpha_1, \ldots, \alpha_{C-1}) \\
&= \begin{cases}
-\sum_{(\tau_i, \tau_j)} \sum_{c=1}^{C-1} \mathbb{I}_{\{y_{(ij)}=c\}} \log \left[G\left\{ \alpha_c - \left(\sum_{s \in \tau_i} r_\theta(s) - \sum_{s \in \tau_j} r_\theta(s) \right) \right\} \right. \\
\qquad \left. -G\left\{ \alpha_{c-1} - \left(\sum_{s \in \tau_i} r_\theta(s) - \sum_{s \in \tau_j} r_\theta(s) \right) \right\} \right] & (y_{ij} = 1, \ldots, C-1) \\
-\sum_{(\tau_i, \tau_j)} \log \left[1 - G\left\{ \alpha_{c-1} - \left(\sum_{s \in \tau_i} r_\theta(s) - \sum_{s \in \tau_j} r_\theta(s) \right) \right\} \right] & (y_{ij} = C)
\end{cases}
\end{aligned}
\tag{3.1}
$$

とする．ここで, 関数 G は $G(x) = \dfrac{\exp(x)}{1+\exp(x)}$ である．

提案手法による報酬関数と方策の学習は以下のように行われる．まず収集した軌跡 $\tau_i, (i = 1, \ldots, m)$ から二つの軌跡 (τ_i, τ_j) をランダムに取り出し, 何らかの基準をもとにカテゴリー数 C の順序尺度のラベルをつける．これを繰り返すことで一対比較データを作成する．作成した一対比較データと損失関数 (3.1) のもとで確率的勾配降下法を用いてパラメータ $\theta, \alpha_1, \ldots, \alpha_{C-1}$ を推定する．そして, 推定したパラメータ $\hat{\theta}$ による報酬関数 $r_{\hat{\theta}}(s)$ のもとで強化学習を実行し, 最適方策 π^* を求める．

4 おわりに

逆強化学習は熟練者の軌跡から報酬関数を推定する教師あり学習であり, ベイズ推定 [3] や最大エントロピー原理 [4] など統計的な手法が用いられている．熟練者の軌跡を収集することが難しいという課題に対して, 一対比較データを用いた逆強化学習である T-REX が提案された．本研究では, T-REX に基づき順序尺度型の一対比較データを用いる手法を提案した．T-REX や提案手法はタスクの目標が変化しないという仮定のもとで報酬を外挿していることに注意が必要である．

今後の課題として, 一対比較データのラベル付けの方法が挙げられる．現実問題では真の報酬が未知であるため, 真のリターンによって軌跡ペアの優劣を評価できない．他の方法として, 後から生成された軌跡ほど優れているとして優劣をつける方法があるが, 誤ったラベルをつけてしまう可能性がある．このようなノイズに対し, 報酬関数の学習にどの程度の影響が出るかについて理論的保障を与えることが今後の課題である．

参考文献

[1] Agresti, A. (1992). Analysis of ordinal paired comparison data, In *Journal of the Royal Statistical Society. Series C (Applied Statistics), 41(2)*, pp.287-297.

[2] Brown, D. S., Goo, W., Prabhat, N., and Niekum, S. (2019). Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations, In *Proceedings of 36th International Conference on Machine Learning*, pp.783-792.

[3] Ramachandran, D., Amir, E. (2007). Bayesian inverse reinforcement learning, In *Proceedings of the 20th international joint conference on Artifical intelligence*, pp.2586-2591.

[4] Ziebart, B. D., Maas, A., Bagnell, J. A. (2008) Maximum entropy inverse reinforcement learning, In *Proceedings of 23th Association for the Advancement of Artificial Intelligence*, pp.1433-1438.