효율적인 다중태스크 오프라인 모델기반 강화학습 알고리즘에 대한 연구

김건형°, 장영수, 이종민, 김기웅
한국과학기술원
{ghkim, ysjang, jmlee}@ai.kaist.ac.kr, kekim@kaist.ac.kr

A Study on Efficient Multi Task Offline Model-based Reinforcement Learning

Geon-Hyeong Kim, Youngsoo Jang, Jongmin Lee, Kee-Eung Kim
KAIST

요 약

오프라인 강화학습은 사전에 수집된 데이터로부터 환경과의 추가적인 상호작용 없이 정책을 학습하는 것을 목표로 한다. 하지만 이러한 프레임워크에서는 단일 태스크에서 사전에 많은 데이터가 수집되어 있어야 한다는 제약이 있으며, 이를 완화하기 위해 다양한 태스크에서 수집된 데이터들을 활용하는 다중태스크 오프라인 강화학습 문제를 생각해 볼 수 있다. 본 논문에서는 이러한 다중태스크 오프라인 강화학습 문제에서 다른 태스크의 데이터를 효율적으로 활용하는 모델기반 강화학습 알고리즘을 제안한다. 제안하는 알고리즘인 MT-OMBRL은 태스크 사이의 동역학 정보를 공유하여 각 태스크를 독립적으로 해결하는 오프라인 강화학습 알고리즘 대비 뛰어난 성능을 보인다.

1 서론

오프라인 강화학습은 환경과의 상호작용 없이 사전에 수집된 데이터만으로 에이전트를 학습하는 방법론으로, 학습중인 에이전트의 정책을 실제로 수행하는 것이 위험한 자율주행, 로보틱스, 헬스케어 등의 현실적인 문제에 적용할 수 있다는 장점이 있다. 하지만 이를 위해서는 풀고자 하는 태스크에서 다수의 데이터가 수집되어 있어야 한다는 제약조건이 있다. 해당 제약조건을 완화하기 위해 소수의 풀고자 하는 태스크의 데이터와 다수의 다양한 태스크의 데이터를 활용하는 다중태스크 오프라인 강화학습 문제를 생각해 볼 수 있다. 이러한 프레임워크는 특정 로봇이 다양한 태스크에서 수집한 데이터셋을 활용한 로봇제어 학습 혹은 다양한 목표지점으로 주행한 데이터셋을 활용한 자율주행 학습 등에 활용할 수 있다. 하지만, 다중태스크 오프라인 강화학습 에이전트 학습에는 다음 두 가지 어려움이 있다.

첫 번째로, 오프라인 강화학습에서 잘 알려진 분포 이동 (distribution shift)문제가 있다. 이는 데이터를 수집한 정책과 현재 학습중인 정책의 분포가 달라지면서 발생하는 문제로, 특히 가치 함수 (value function)를 부트스트랩 기법 (bootstrapping)을 통해 업데이트 하는 경우 가치함수를 과대평가하게 된다. 이러한 과대평가된 가치함수는 실제로는 좋지 않은 행동을 과대평가하게 되고, 이를 통한 정책의 학습은 에이전트의 성능저하를 야기한다. 이를 방지하기 위해 대다수의 오프라인 강화학습 연구들은 보수적인 가치함수 학습을 통해 과대평가를 방지한다 [1, 2].

두 번째로, 다양한 태스크에서 수집된 데이터들을 단순한 방식으로 공유하여 활용할 경우 오히려 풀고자 하는 태스크에서 수집된 데이터셋만 활용하는 방법에 비해 성능이 저하될 수도 있다는 것이다 [3]. 이에 기존 연구들에서는 환경과의 상호작용이 가능한 온라인 상황에서 다른 태스크에서 수집된 데이터들을 효과적으로 활용하기 위한 다양한 방법들이 제시되었다 [3, 4]. 하지만, 오프라인 상황에서는 아직 충분한 연구가 진행되지 않았으며, 최근 진행된 다중태스크 오프라인 강화학습 연구는 보상함수에 대한 접근이 가능하다는 강한 가정 하에서 진행되었다 [5].

본 연구에서는 앞선 연구에서의 가정을 완화하기 위해 보상함수에 대한 접근이 불가능한 경우에서의 다중태스크 오프라인 강화학습 문제를 다루고자 한다. 이를 위해 다양한 태스크에서 생성된 모든 데이터를 추가적인 보상함수 재할당 없이 활용하여 모델을 학습하는 한편, 오프라인 강화학습에서 쓰이는 보수적인 가치함수 학습을 활용하는 방안을 제시한다. 이를 통해 단순히 풀고자 하는 태스크의 데이터만 활용하는 오프라인 강화학습 알고리즘 대비 뛰어난 성능을 보이는 다중태스크 오프라인 강화학습 알고리즘을 제안한다.[1]

2 연구 배경

본 연구에서는 다중태스크 마르코프 의사결정과정 (multi-task MDP; [5]) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, \gamma, \{R_i, i\}_{i=1}^{N} \rangle$의 형태로 정의할 수 있는 다중태스크 문제들을 고려한다. $\mathcal{S}$는 에이전트가 직면할 수 있는 상태의 집합, $\mathcal{A}$는 에이전트가 할 수 있는 행동의 집합, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$는 환경의 동역학을 표현하는 상태 전이 함수, $\gamma \in [0, 1)$는 시간에 따른 보상의 중요도의 감소 정도를 나타내는 감쇠 상수, $R_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$는 각 태스크 $i \in [N]$ ($[N] = \{1, 2, \ldots, N\}$을 나타낸다)에서 상태와 행동에 따라 제공되는 보상 함수이다. 또한, 각 다중태스크 MDP는 보상 함수만 다르고 동일한 동역학 구조를 공유한다고 가정한다. 이러한 환경 하에서 에이전트의 정책은 각 태스크별 확률적 매핑으로, $\pi : \mathcal{S} \times [N] \to \Delta(\mathcal{A})$으로 정의되며, 편의상 태스크 i에서 상태 s가 주어졌을 때 행동 a에 대한 매핑의

1) 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-00940, 안전한 강화학습 원천 기술 개발 및 자연어 처리에의 응용, 및 No.2017-0-01779, 의사결정 이유를 설명할 수 있는 인간 수준의 학습·추론 프레임워크 개발)

값을 $\pi(a|s,i)$로 표기한다.

다중태스크 오프라인 강화학습은 각각의 태스크 i에서 사전에 수집된 데이터셋 $\mathcal{D}_i = \{(s_j, a_j, r_j, s'_j)\}_{j=1}^M$ 을 활용하여 환경과의 추가적인 상호작용 없이 보상합의 기댓값 $\mathbb{E}_{i\sim[N]}\mathbb{E}_{\pi(\cdot|\cdot,i)}[\sum_t \gamma^t R_i(s_t,a_t)]$을 최대화하는 최적정책 π^*를 찾는 문제이다. 가장 쉬운 방법은 각 태스크 i에 대한 정책 $\pi(a|s,i)$를 $\mathcal{D}_i$만 활용하여 오프라인 강화학습 알고리즘으로 학습하는 것이다. 하지만, 다른 태스크에서 수집한 데이터 또한 동역학 구조에 대한 정보는 가지고 있기 때문에 이를 활용하여 구조에 대해 보다 정확히 학습하는 것이 성능 향상에 도움이 될 여지가 충분하다. 기존 다중태스크 오프라인 강화학습 연구에서는 보상함수 R_i의 함수 형태를 모두 안다는 가정 하에서 태스크 i에 대한 정책 학습 시, 타 태스크 j ($i \neq j$)의 보상값을 현재 태스크의 보상함수를 통해 재할당하여 활용한다 [5]. 하지만 이러한 가정은 일반적으로 항상 성립하지 않는 강한 가정으로, 본 연구에서는 이를 완화하기 위해 보상함수에 접근하지 못하는 상황에 대해 다루고자 한다.

3 본론

모델기반 강화학습은 일반적으로 최대 가능도 방법 (maximum likelihood estimation; MLE) 등을 활용하여 전이 함수 P와 보상함수 R을 잘 근사하는 함수 $\hat{P}$와 $\hat{R}$을 찾고, 이를 기반으로 최적 정책을 찾는 형태로 구성된다. 다중태스크 강화학습은 각 태스크가 동일한 전이 함수와 서로 다른 보상 함수를 갖기 때문에 모델기반 다중태스크 강화학습 에이전트는 태스크 공통의 전이 함수 $\hat{P}$와 각 태스크별 보상 함수의 집합 $\hat{\mathbf{R}} = \{\hat{R}_i\}_{i=1}^N$를 적절한 방법으로 학습해야 한다. 이후 학습한 다중태스크 모델에 기반하여 최적 정책을 학습하는데, 이 때 오프라인 강화학습에서 널리 쓰이는 보수적인 학습 방법을 다중태스크 상황에 맞도록 적절하게 적용하는 것이 중요하다.

본 연구에서는 다양한 태스크에서 수집된 데이터를 효율적으로 활용하기 위해 전체 데이터셋 $\mathcal{D} = \cup_{i=1}^N \mathcal{D}_i$를 활용하여 전이 함수 P를 학습하고, 각 태스크 i에 대해 해당 태스크의 $\mathcal{D}_i$만을 활용하여 보상 함수 R_i를 학습한다. 이러한 학습 방법은 전이 함수와 보상함수가 각각 활용할 수 있는 최대한의 데이터를 활용하는 방법이기 때문에 직관적으로 적절한 학습 방법이라고 할 수 있다. 각각의 함수는 MLE를 사용하여 학습하며, 유한한 상태 및 행동을 갖는 유한 MDP (finite MDP)의 경우 다음 방법으로 계산할 수 있다:

$$\hat{P}(s,a,s') = \frac{n(s,a,s')}{n(s,a)}, \qquad \hat{R}_i = \frac{\bar{r}_i(s,a)}{n_i(s,a)}. \tag{1}$$

여기서, $n(s,a)$와 $n(s,a,s')$는 각각 전체 데이터셋 $\mathcal{D}$ 내에 있는 (s,a)와 (s,a,s')의 개수이며, $\bar{r}_i(s,a)$와 $n_i(s,a)$는 각각 태스크 i에서 수집된 데이터셋 $\mathcal{D}_i$ 내에 있는 $r(s,a)$의 평균값과 (s,a)의 개수를 의미한다.

Algorithm 1 MT-OMBRL

Require: 각 태스크 i에서 사전 수집된 데이터 $\{\mathcal{D}_i\}_{i=1}^N$, 상수 $\beta \geq 0$

1: 근사 전이 함수 $\hat{P}$ 계산 (수식 (1))
2: **for** $i = 1$ to N **do**
3: 근사 보상 함수 $\hat{R}_i$ 계산 (수식 (1))
4: 전이 함수 $\hat{P}$와 페널티를 반영한 보상 함수 $\hat{R}_i + e_i$ (수식 (2))에 대해 DP를 통해 최적 정책 $\pi(\cdot|\cdot,i)$ 계산
5: **end for**

모델에 기반한 정책 학습을 할 때, 온라인 모델기반 강화학습의 경우 불확실성에 기반한 보너스를 제공하는 불확실성에 대한 낙관론적 (optimism in the face of uncertainty) 접근법이 널리 쓰이는 반면에 오프라인 모델기반 강화학습의 경우 불확실성에 기반한 페널티를 제공하는 불확실성에 대한 비관론적 (pessimism in the face of uncertainty) 접근법이 일반적이다. 이 때, 유한 MDP 문제에 대해 온라인 모델기반 강화학습은 상태-행동 쌍 (state-action pair) (s,a)의 관측 횟수 $n(s,a)$와 임의의 양의 상수 $\beta \in \mathbb{R}_{\geq 0}$에 기반한 보너스 $\frac{\beta}{\sqrt{n(s,a)}}$를 학습한 보상함수 $\hat{R}(s,a)$에 더하는 방법이 잘 알려져 있다 [6]. 이와는 반대로, 오프라인 모델기반 강화학습은 $-\frac{\beta}{\sqrt{n(s,a)}}$를 학습한 보상함수 $\hat{R}(s,a)$에 대한 페널티로 더해주는 방법을 생각해 볼 수 있다. 이러한 불확실성 패널티에 기반한 오프라인 강화학습 알고리즘은 단일태스크 환경일 때 높은 확률로 데이터수집 정책 대비 성능 향상을 달성할 수 있음이 이론적으로 증명된 바 있다 [7]. 하지만, 멀티태스크 환경에서 기존 모델기반 강화학습 알고리즘들은 전이 함수와 보상 함수가 같은 불확실성을 공유하는 반면에 앞서 학습한 전이 함수 $\hat{P}$와 보상 함수의 집합 $\hat{\mathbf{R}}$은 서로 다른 양의 데이터를 사용하여 서로 다른 불확실성을 갖는다. 즉, 태스크 i에 대한 정책 학습 시 페널티를 $n(s,a)$와 $n_i(s,a)$ 중 어떤 값을 기반으로 할지에 대해 명확하지 않다. 본 연구에서는 두 가지를 모두 고려한

$$e_i(s,a;\beta) := -\beta\left(\frac{1}{\sqrt{n(s,a)}} + \frac{1}{\sqrt{n_i(s,a)}}\right) \tag{2}$$

를 페널티로 사용하여 멀티태스크 오프라인 모델기반 강화학습 (multi-task offline model-based reinforcement learning; MT-OMBRL)을 제안한다. 해당 방법론은 근사 전이 함수 $\hat{P}$와 페널티가 반영된 근사 보상 함수 $\hat{R}_i + e_i$에 대해 벨만 최적 방정식 (Bellman optimality equation)을 활용한 동적 계획법 (dynamic programming; DP)으로 정책을 계산한다 [8]. 사용한 최종 알고리즘의 의사 코드는 Algorithm 1에 나타내었다.

4 실험

제안하는 알고리즘의 성능을 평가하기 위해 유한 상태 및 행동 공간을 갖는 문제 중 하나인 Random MDP 도메인 [9]에서 실험을 진행하였다. 태스크의 개수는 5개이며, 각 태스크는 20개의 상태

와 4개의 행동을 가지며, 초기 상태는 1로 고정되어 있다. 각 태스트마다 (s, a) 쌍은 임의의 서로 다른 4개의 상태로 전이할 수 있으며, 이 때의 전이 확률은 $\mathrm{Dirichlet}(1, 1, 1, 1)$ 분포에서 샘플링 된다. 전이함수는 태스크 종류에 관계없이 모두 공유된다. 보상함수는 각 (s, a)마다 파라미터 $p_{s,a}$의 베르누이 분포를 따르며 $p_{s,a}$는 $\mathrm{Beta}(1, 1)$에서 샘플링 된다. 보상함수는 태스트 종류마다 서로 다르게 정의된다. 감쇠상수는 0.95로 설정되었다. 이러한 도메인 하에서 우리는 각각의 태스크에 대해 최적 정도 (degree of optimality)를 다양하게 조절하여 데이터를 수집하였으며, 이러한 최적 정도는 기존 Random MDP를 사용한 연구를 따른다 [9]. 데이터 수집 정책의 최적 정도가 낮을 수록 uniform policy에 가까워지므로 수집된 오프라인 데이터셋이 넓은 상태-행동 영역을 커버하게 된다. 실험 결과는 데이터 수집 정책의 성능을 0, 최적 정책의 성능을 1로 정규화한 성능을 보여준다.

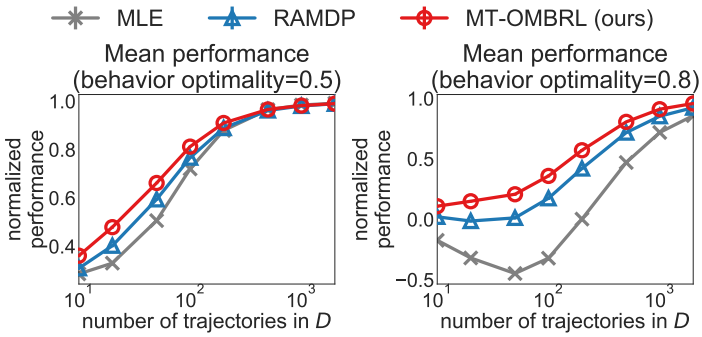

그림 1: Random MDP에서 10000번 반복 실험 수행 결과

그림 1에서는 제안한 MT-OMBRL를 두 가지 베이스라인과 비교하였다. 첫 번째로, 각 태스크별 독립적으로 학습한 오프라인 모델기반 알고리즘 RAMDP를 하나의 베이스라인으로 사용하였다. 해당 알고리즘은 MT-OMBRL이 태스크별 오프라인 강화학습 알고리즘 대비 더 효율적으로 데이터를 사용하는 것을 보여주기 위해 설계되었으며, 풀고자 하는 태스크에서 수집한 데이터만 사용해 전이 함수와 보상 함수를 MLE로 근사한 후 페널티 $\frac{\beta}{n_i(s,a)}$를 보상함수에 반영하여 DP를 수행하였다. 두 번째로, MT-OMBRL와 같은 방법으로 MLE를 활용해 전이 함수와 보상 함수를 근사하되, 보상함수에 페널티를 반영하지 않은 알고리즘을 다른 하나의 베이스라인으로 사용하였다. 해당 알고리즘은 본 연구에서 제안한 페널티의 효용성을 보이기 위한 것으로, 그림 1에서 MLE라고 표기하였다. 실험 결과를 통해 본 연구에서 제안한 MT-OMBRL이 각 태스크에 독립적으로 학습한 오프라인 모델기반 알고리즘 RAMDP 대비 뛰어난 성능을 보임을 확인할 수 있었다. 또한, 단순 MLE로만 학습하는 경우 전반적으로 오프라인 상황에서 성능이 떨어지는 것을 확인하였고, 특히 그림 1의 우측과 같이 데이터가 늘어남에도 불구하고 오히려 성능이 저하하기도 하는 점을 확인하였다. 이를 통해 다중태스크 오프라인 강화학습 문제에서도 오프라인 강화학습과 마찬가지로 불확실성에 대한 페널티가 중요함을 확인할 수 있었다.

5 결론

본 논문에서는 효율적인 다중태스크 오프라인 강화학습 알고리즘을 위해 모델기반의 알고리즘 MT-OMBRL을 제시하였다. 제안하는 방법론은 다양한 태스크에서 수집된 데이터를 효율적으로 활용하기 위해 다중태스크 오프라인 상황에 적절한 모델을 학습 및 불확실성에 대한 페널티를 디자인하였다. 이를 통해 각 태스크를 독립적으로 학습하는 것에 비해 성능이 향상됨을 보여준다. 추후 연구 방향으로는 로봇 제어 시뮬레이션 환경과 같이 연속 MDP (continuous MDP)로 표현되는 문제들에도 적용이 가능하도록 방법론을 확장 및 추가 실험 진행 등을 고려할 수 있다.

참고 문헌

[1] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *International Conference on Machine Learning (ICML)*, 2019.

[2] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine, "Stabilizing off-policy Q-learning via bootstrapping error reduction," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[3] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman, "Mt-opt: Continuous multi-task robotic reinforcement learning at scale," in *Conference on Robot Learning (CoRL)*, 2021.

[4] B. Eysenbach, X. Geng, S. Levine, and R. Salakhutdinov, "Rewriting history with inverse rl: Hindsight inference for policy improvement," *arXiv preprint arXiv:2002.11089*, 2020.

[5] T. Yu, A. Kumar, Y. Chebotar, K. Hausman, S. Levine, and C. Finn, "Conservative data sharing for multi-task offline reinforcement learning," *arXiv preprint arXiv:2109.08128*, 2021.

[6] A. L. Strehl and M. L. Littman, "An analysis of model-based interval estimation for markov decision processes," *Journal of Computer and System Sciences*, vol. 74, 2008.

[7] M. Petrik, M. Ghavamzadeh, and Y. Chow, "Safe policy improvement by minimizing robust baseline regret," in *Advances in Neural Information Processing Systems*, 2016.

[8] R. S. Sutton, A. G. Barto, *et al.*, *Introduction to reinforcement learning*, vol. 135. MIT press Cambridge, 1998.

[9] R. Laroche, P. Trichelair, and R. T. Des Combes, "Safe policy improvement with baseline bootstrapping," in *International Conference on Machine Learning*, 2019.