복잡계 네트워크를 이용한 강화 학습 구현

이승준* 장병탁
서울대학교 비오지능연구실
{sjlee*, btzhang}@bi.snu.ac.kr

Reinforcement Learning with Small World Network

Seungjoon Yi† Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

강화 학습(Reinforcement Learning)을 실제 문제에 적용하는 데 있어 가장 큰 문제가 차원성의 저주(Curse of dimensionality)이다. 문제가 커지면 따라 목적을 이루기 위해서 더 많은 단계의 판단이 필요하고 이에 따라 문제의 해결이 지속적으로 어려워지게 된다. 이를 해결하기 위해 문제를 여러 단계로 나누어 단계별로 학습하는 계층적 강화 학습(Hierarchical Reinforcement Learning)이 제시된 바 있다. 하지만 대부분의 계층적 강화 학습 방법들은 각 단계에 문제가 구조를 가진 것을 설계하여 전제에 큰 사이즈의 문제를 간단히 표현할 방법을 제시하지 않는다는. 따라서 이를 방법들은 실제적인 문제에 바로 적용하기에는 적합하지 않다. 이러한 문제점들을 해결하기 위해 복잡계 네트워크(Complex Network)가 갖는 작은 세상 성질(Small world property)에 의하여 자기조직화하는 생장 네트워크(Self organizing growing network)를 기반으로 한 합리적 표현 모델이 제안된 바 있다. 이러한 모델에서는 문제 크기가 커지더라도 네트워크의 사이즈가 크게 커지지 않기 때문에 문제의 난이도가 크게 증가하지 않을 것을 기대할 수 있다. 본 논문에서는 이러한 합리적 표현 모델을 사용한 강화 학습 알고리즘을 구현하고 실험을 통해 각 모델이 강화 학습의 문제 사이즈에 따른 성능에 미치는 영향에 대해 알아보고자 한다.

1. 서 론

강화 학습(Reinforcement Learning)에서는 에이전트는 환경(World)과 상호작용하며 최대의 보상(Reward)을 주는 상태(State)와 행동(Action)의 샘플을 정책(Policy)을 학습하여 한다. 전통적인 RL 프레임워크에서는 환경이 이산적인 상태 및 공간으로 이루어진 마르코프 결정 프로세스(Markov Decision Process)로 정의되며, Q-Learning과 같은 강화 학습 알고리즘에서는 상태와 행동 공간을 데이블의 형태로 가정하고 모든 상태-행동의 평가값(Value function)을 구해서 최적의 정책을 결정하는 것이다. 하지만 대부분의 실용적 문제는 상태의 차원이 매우 많거나 연속적인 상태를 가지는 경우가 많고, 이런 경우 모든 상태-행동의 평가값을 구하는 것이 불가능하다. 상태가 이산적인 경우에도 문제가 커지면 모든 상태-행동의 평가값을 구하는 것은 현실적으로 어려워지게 된다. 따라서 신경망과 같은 함수 근사장치를 강화 학습에 사용하는 방식이 주로 사용되어 왔다. 하지만, 함수 근사장치를 사용하더라도 문제가 커지면 따라 학습해야 할 파라미터의 수가 지속적으로 증가하게 되고 결국 복잡성을 높임으로써 차원성의 저주를 피할 수 없다 [2].

이 차원성의 저주는 행동의 단계마다 판단을 해야 하기 때문에 나타난다. 따라서 이를 해결하기 위한 방법으로 많은 단계의 판단을 하기 위한 방법은 제안되었고, 나아가 계층적인 차원 구조와 이에 따른 학습 방법인 계층적 강화 학습이 제안되었다. 하지만 대부분의 계층적 강화 학습 알고리즘은 두 가지의 문제를 가지고 있다. 문제의 계층적 구조를 사전에 미리 알야 하고, 상태와 행동 공간을 여전히 데이블의 형태로 가정하는 것이다 [2]. 즉 계층적 강화 학습 알고리즘도 실제 문제에 적절 적용하기에는 한계가 있다. 본 논문은 최근의 복잡계 네트워크에 대한 연구 결과와 함께 그레프, 사회 네트워크 등의 많은 실제의 네트워크들이 여러 공통된 성질을 만드는 것이 알려졌다. 그 중 하나는 ‘작은 세상 성질(Small World Property)’인데, 이는 비교적 작은 사이즈의 네트워크라도 대부분의 노드들 사이에 짧은 경로가 존재한다는 것이다. 또한, 이러한 짧은 경로를 찾을 수 있는 비 동정적적인 탐색 방법이 존재한다는 것도 알려져 있다 [3].

이 성질을 이용하여 복잡계 네트워크 형태의 환경 모델을 사용한 학습 방법이 제시되었다 [8]. 네트워크가 작고 많은 성질을 따르며 유지하면 문제가 커지더라도 판단 단계의 수가 크게 늘어나지 않게 할 수 있으므로 차원성의 저주를 피할 수 있다. 본 논문에서는 이러한 성질을 사용하여 실제로 강화 학습을 구현하고, 실험을 통해 강화 학습의 문제 사이즈에 따른 성능에 대해 알아보고자 한다.

2. 관련 연구

2.1. 복잡계 네트워크

복잡계 네트워크는 대부분의 실제 세계의 네트워크들은 작은 세상 성질(Small world property), 높은 클러스터 계수(Compact cluster coefficient), 적도 없는 도수 분포(Scale–free degree
distribution)과 같은 세 가지 성질을 가진다 [4].


사람의 경우 개인이 전체 네트워크에 대해 모르더라도 어느 경로가 더 가능성이 있는지 판단함으로써 좋은 경로를 찾아낼 수 있다. 이 사실에 기반해서 복잡성 네트워크에서 효율적인 비 중점지점 탐색 알고리즘을 연구하였다. [7]의 모델에서는 이러한 좋은 경로를 찾는 시간에 찾아내는 비 중점지점 알고리즘을 존재할 수 있다는 것이 증명되어 있다. 하지만 이 모델의 내부 구조를 탐색에 사용할 수 있도록 수용한 것이 필요하다 [3].

내부 구조 탐색에 전부 사용할 수 없을 경우에도 네트워크의 적도 없는 도수 분포를 가진다면 효율적인 탐색이 가능하다는 것도 알려져 있다.

2.2 복잡한 네트워크를 사용한 환경 모델

[8]에서 제안한 환경 모델에서 자기 조직화하는 성질

1. 행동 a를 행하고 다음 상태 x’와 보상 r을 받는다.
2. ITPM에서 x’에 가장 가까운 노드 b’를 찾는다.
3. x’가 b’에서 멀리 떨어져 있을 경우 새로운 노드 b’의 위치에 결정하고 5번으로 간다.
4. b’의 Q를 사용해서 다음 행동 a’를 선택한다.
5. RL 알고리즘을 사용해서 기존의 가장 가까운 노드 b의 Q 값을 수집한다.
6. 자기조직화된 b’의 연결 상태와 위치를 수정한다.

작은 세밀성 둘 다가 하기 위해 다음과 같은 자기조직화 알고리즘을 사용한다.[8]

\[(b-1)\text{의 노드 } n\text{의 확률분포에 따라 선택한다.}
\]

\[
\begin{align*}
\text{MODEL 1:} \quad & d(u, v) = r \\
\text{MODEL 2:} \quad & d(v) \\
\end{align*}
\]

\[(b-1)\text{의 } u\text{와 } v\text{를 연결한다.}
\]

결과적으로 생성되는 네트워크는 ITPM이 생성하는 균일한 격자 구조의 네트워크 위에 작은 세밀하게 의한 정격리 링크가 추가된 형태가 된다.

3. 복잡한 네트워크를 사용한 강화 학습

3.1 행동 선택

Q-Learning과 같은 대부분의 강화 학습 알고리즘의 경우 행동 선택 방법에는 큰 제약이 주어지지 않는다. 실제로 행동 선택 방법과 상관없이 모든 상태가 충분히 회수 이상 방문될 경우 최적의 해를 구할 수 있다.


3.2 Q값 업데이트

온라인의 네트워크에서 링크가 추가된 네트워크는 계층적 강화 학습의 보완된 MDP의 대응으로, 계층적 강화 학습의 Q값 업데이트 알고리즘을 사용할 수 있다. [6]에서 말한 각 링크가 수행 시간 $k(s, o)$를 가지고 링크를 따르는 경우 보상 $r$이 아니라 링크들의 평균값 $\bar{r}$과 같이 수행 가능하다. 학습률의 파라미터 $\eta$의 경우 원활한 수렴을 위해 학습 진행시에 따라 점점 줄여도 될 수 있다.

\[Q(s, a) = Q(s, a) + \alpha [r + \gamma Q(s', a) - Q(s, a)] \quad (1)\]

4. 실험 및 결과

제안된 두 가지 모델과 결합된 적용된 ITPM 알고리즘은 사용하여 2차원 상의 문제에 적용하여 보았다. ITPM에는 $\rho = 0.0009, \delta = 0.0002, \theta = 0.0002, \psi = 2.322, \mu_{\text{init}} = 0.2$의 파라미터들이 사용되었고 Q-Learning에는 $\varepsilon = 0.1, \alpha = 0.5, \gamma = 0.7$의 파라미터들이 사용되었다. 보상 $r$은 Goal에서 10, 나머지의 경우 1을 사용하였다.

4.1 생성된 모델 비교

성실한 경향을 얻은 환경 하에서 각 알고리즘을 사용하여 다양한 해상도의 네트워크를 학습하였다. 학습된 그래프의 모양과 도수 분포는 그림 1과 같다.

![그림 1. 각 모델별 학습된 그래프](image-url)
문제 사이즈에 따른 평균 노드간 거리의 변화는 그림 2와 같았다. 복잡한 네트워크 모델을 사용하지 않은 경우 노드간 거리가 문제 사이즈에 비례해서 증가하지만 사용한 경우 문제 사이즈의 log값에 비례해서 증가함을 알 수 있다.

![그림 2. 노드간 평균 거리의 변화](image)

4.2 강화 학습 성능 비교
세 모델을 사용하여 Q-learning 알고리즘을 수행하여 학습 성능을 비교하여 보았다. 그림 3에 문제 사이즈에 따른 강화 학습의 학습 곡선이 나타난다. 사이즈가 커짐에 따라 복잡한 네트워크 모델을 사용할 경우 수렴속도가 크게 빨라짐을 알 수 있다.

![그림 3. 문제 사이즈에 따른 학습 곡선](image)

평균 학습 시간을 최종 수렴값의 절반에 다다르기까지의 세대수로 정의하였다. 그림 4는 문제 사이즈에 따른 평균 학습 시간을 나타낸다. 복잡한 네트워크 모델을 사용하지 않은 경우보다 사용한 경우 학습 시간의 증가폭이 현저하게 줄어들었음을 알 수 있다. 모델 1과 모델 2를 비교해 보면 문제 사이즈에 따른 노드간 평균 거리의 변화의 양상은 거의 유사하지만 강화 학습의 성능은 모델 1이 크게 떨어지는데, 이는 모델 2가 적도 없는 도수 분포를 가정으로서 병도의 탐색 알고리즘 없이 효율적인 탐색이 가능해지기 때문이라고 생각된다.

![그림 4. 평균 학습 시간의 변화](image)

감사의 글
이 논문을 교육인적성태서 BK11 사업과 산업지원부에 의해 지원되었음.

참고 문헌
[8] 이승준, 장병탁. 복잡한 네트워크를 이용한 강화 학습에서의 환경 표현. 한국정보과학회 부 학술발표 논문 집 (B), 제 31권 1호, pp.622-624,2004