복잡계 네트워크를 이용한 강화 학습 구현

이승준° 장병탁
서울대학교 바이오지능연구실
{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 문제는 차원성의 저주(Curse of dimensionality)이다. 문제가 커짐에 따라 목적을 이루기 위해서 더 많은 단계의 판단이 필요하고 이에 따라 문제의 해결이 지수적으로 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 사전에 문제의 구조를 아는 것을 전제로 하며 큰 사이즈의 문제를 간단히 표현할 방법을 제시하지 않는다. 따라서 이들 방법들도 실제적인 문제에 바로 적용하기에는 적합하지 않다.

이러한 문제점들을 해결하기 위해 복잡계 네트워크(Complex Network)가 갖는 작은 세상 성질(Small world property)에 착안하여 자기조직화하는 생장 네트워크(Self organizing growing network) 를 기반으로 한 환경 표현 모델이 제안된 바 있다. 이러한 모델에서는 문제 크기가 커지더라도 네트워크의 사이즈가 크게 커지지 않기 때문에 문제의 난이도가 크기에 따라 크게 증가하지 않을 것을 기대할 수 있다.

본 논문에서는 이러한 환경 모델을 사용한 강화 학습 알고리즘을 구현하고 실험을 통하여 각 모델이 강화 학습의 문제 사이즈에 따른 성능에 끼치는 영향에 대해 알아보았다.

1. 서 론

강화 학습 (Reinforcement Learning) 에서는 에이전트는 환경(World)과 상호작용하며 최대의 보상(Reward)를 주는 상태(State)와 행동(Action)의 함수인 정책(Policy)를 학습하려 한다. 전통적인 RL 프레임웍에서는 환경을 이산적인 시간과 공간으로 이루어진 마르코프 결정 프로세스(Markov Decision Process)으로 정의하고, Q-Learning 과 같은 강화 학습 알고리즘에서는 상태와 행동 공간을 테이블의 형태로 가정하고 모든 상태-행동의 평가값(Value function)을 구해서 최적의 정책을 결정하게 된다.

하지만 대부분의 실세계 문제는 상태의 차원이 매우 많거나 연속적인 상태를 가지는 경우가 많고, 이런 경우 모든 상태-행동의 평가값을 구하는 것이 불가능하다. 상태가 이산적인 경우에도 문제가 커짐에 따라 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 커짐에 따라 학습해야 할 파라미터의 수가 지수적으로 증가하게 되고, 결국 복잡도를 높임에 따라 나타나는 차원성의 저주는 피할 수 없다 [2].

이 차원성의 저주는 행동의 단계마다 판단을 해야 하기 때문에 나타난다. 따라서 이를 피하기 위한 방법으로 한번의 판단으로 여러 행동을 하게 하는 방식이 제안되었고, 나아가 계층적인 제어 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 알아야 하고, 상태와 행동 공간을 여전히 테이블의 형태로 가정하는 것이다 [2]. 즉 계층적 강화 학습 알고리즘도 실제 문제에 직접 적용하기에는 한계를 가진다.

한편 최근의 복잡계 네트워크에 대한 연구 결과 웹 그래프, 사회 네트워크 등의 많은 실세계의 네트워크들이 여러 공통된 성질을 띤다는 것이 알려졌다. 그 중 하나가 '작은 세상 성질(Small World Property)' 인데, 이는 비교적 큰 사이즈의 네트워크라도 대부분의 노드들 사이에 짧은 경로가 존재한다는 것이다. 또한, 이러한 짧은 경로를 찾아낼 수 있는 비 중앙집중적인 탐색 방법이 존재한다는 것도 알려져 있다 [3].

이 성질을 이용하여 복잡계 네트워크 형태의 환경 모델을 사용한 환경 표현 방법이 제시되었다 [8]. 네트워크가 작은 세상 성질을 띠도록 유지하면 문제가 커지더라도 판단 단계의 수가 크게 늘어나지 않게 할 수 있으므로 차원성의 저주를 피할 수 있다.

본 논문에서는 이러한 환경 모델을 사용하여 실제로 강화 학습을 구현하고, 실험을 통해 강화 학습의 문제 사이즈에 따른 성능에 각 모델이 끼치는 영향에 대해 알아보았다.

2. 관련 연구

2.1. 복잡계 네트워크

대부분의 실세계 네트워크들은 작은 세상 성질(Small world property), 높은 클러스터 계수(High cluster coefficient), 척도 없는 도수 분포(Scale-free degree

distribution)와 같은 세 가지 성질을 가진다 [4].

많은 실세계 네트워크들이 위의 성질들을 보이고 있기 때문에 이러한 성질을 가지는 네트워크를 모델링하려는 많은 시도가 있어 왔다. 처음 제안된 방법인 [7]에서는 각 노드들이 이웃하고만 연결되어 있는 바둑판 모양의 격자에서 출발해서 임의의 확률로 링크를 추가한다. 생성되는 모델은 작은 세상 성질과 높은 클러스터 계수를 가지게 된다. 척도 없는 도수 분포를 가진 작은 세상 네트워크를 모델링하기 위해 [1]에서는 부익부 빈익빈 모델을 사용한 생장 네트워크 모델을 사용하였다. 한편 세 성질 모두를 가지는 네트워크 모델이 [4]에서 제안된 바 있다.

사람의 경우 개인이 전체 네트워크에 대해 모르더라도 어느 경로가 더 가능성이 있는지 판단함으로써 짧은 경로를 찾아낼 수 있다. 이 사실에 기인해서 복잡계 네트워크에서 효율적인 비 중앙집중적 탐색 알고리즘이 연구되어 왔다. [7]의 모델에서는 이러한 짧은 경로를 빠른 시간에 찾아내는 비 중앙집중적 알고리즘이 존재할 수 없다는 것이 증명되어 있다. 하지만 이 모델의 내부 구조를 탐색에 사용할 수 있도록 수정하면 그러한 알고리즘이 가능하다 [3].

내부 구조를 탐색에 전혀 사용할 수 없을 경우라도 네트워크가 척도 없는 도수 분포를 가진다면 효율적인 탐색이 가능하다는 것도 알려져 있다.

2.2 복잡계 네트워크를 사용한 환경 모델

[8]에서 제안된 환경 모델에서는 자기 조직화하는 성장 신경망인 ITPM[5]을 사용하여 환경을 근사하여 네트워크의 형태로 나타낸다. ITPM이 하는 일은 다음과 같다.

1. 행동 a를 행하고 다음 상태 x'와 보상 z를 받는다.
2. ITPM에서 x'에 가장 가까운 노드 b'를 찾는다.
3. x'가 b'에서 멀리 떨어져 있을 경우 새로운 노드를 그 위치에 생성하고 5번으로 간다.
4. b'의 Q값을 사용해서 다음의 행동 a'를 선택한다.
5. RL 알고리즘을 사용해서 기존의 가장 가깝던 노드 b의 Q값을 수정한다.
6. 자기조직화:b'의 연결 상태와 위치를 수정한다.

작은 세상 성질을 띠게 하기 위해 다음과 같은 자기조직화 알고리즘을 사용한다 [8].

(b-ii)노드 v를 다음의 확률분포에 따라 선택한다.
MODEL 1: $distance(u,v)^{-p}$
MODEL 2: $d(v)$
(b-iii) u와 v를 연결한다.

결과적으로 생성되는 네트워크는 ITPM이 생성하는 균일한 격자 구조의 네트워크 위에 작은 세상 모델에 의한 장거리 링크가 추가된 형태가 된다.

3. 복잡계 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 충분한 회수 이상 방문될 경우 최적의 해를 구할 수 있다.

일반적으로 강화 학습의 행동 선택에 많이 쓰이는 방식으로는 일정 확률로 무작위 행동을 하고 나머지의 경우 알려진 것 중 가장 좋은 것을 찾는 ε-greedy 방식이 있다.

하지만 네트워크의 구조에 따라 적절한 비 중앙집중적 탐색 알고리즘을 사용하여 탐색의 효율을 한층 더 높일 수도 있다. [8]의 첫 번째 모델의 경우 노드의 거리를 행동 선택에 이용할 경우 효율적인 탐색이 가능함이 알려져 있다. 반면 척도 없는 도수 분포를 가지는 [8]의 두 번째 모델의 경우 무작위로 이동하더라도 더 높은 차수로 확률적으로 이동하게 되고, 따라서 별도의 탐색 알고리즘을 사용하지 않아도 효율적인 탐색이 가능해짐이 알려져 있다.

실험에서는 ε-greedy 방식을 기반으로 하고 보다 균일한 탐색을 가능하게 하도록 가본 회수에 따라 탐색 가중치를 부여하고 한번 간 node를 다시 가지 않는 taboo search 방식을 결합해서 사용하였다.

3.2 Q값 업데이트

원래의 네트워크에 링크가 추가된 네트워크는 계층적 강화 학습의 보완된 MDP와 대응되므로, 계층적 강화 학습의 Q값 업데이트 알고리즘을 사용할 수 있다. [6]에 따르면 각 링크 가 수행 시간 $k(s,o)$를 가지고 링크를 따를 경우 보상이 r이라면 각 링크들의 평가치를 다음과 같이 수정 가능하다. 학습률인 파라미터 α의 경우 원활한 수렴을 위해 학습 진행에 따라 점점 줄어들도록 할 수 있다.

$Q(s,o)\leftarrow$

$$Q(s,o)+\alpha[r+\gamma^{k(s,o)}max_{o'}Q(s',o')-Q(s,o)] \quad (1)$$

4. 실험 및 결과

제안된 두 가지 모델과 결합된 개량된 ITPM 알고리즘을 사용하여 2차원 상의 문제에 적용하여 보았다. ITPM에는 $\gamma^2=0.00009, \delta=0.0002, \delta_r=0.00002, p=2.322, p_{longlink}=0.2$ 의 파라미터들이 사용되었고, Q-Learning에는 $\varepsilon=0.1, \alpha=0.5, \gamma=0.7$의 파라미터들이 사용되었다. 보상 r은 Goal에서 10, 나머지의 경우 -1을 사용하였다.

4.1 생성된 모델 비교

정사각형의 연속된 환경 하에서 각 알고리즘을 사용하여 다양한 해상도로 네트워크를 학습하였다. 학습된 그래프의 모양과 도수 분포는 그림 1과 같다.

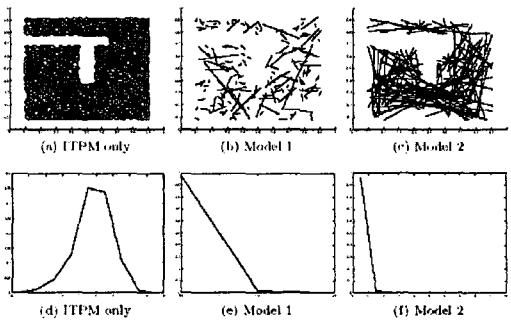

그림 1. 각 모델별 학습된 그래프

문제 사이즈에 따른 평균 노드간 거리는 그림 2와 같다. 복잡계 네트워크 모델을 사용하지 않은 경우 노드간 거리가 문제 사이즈에 비례해서 증가하나 사용한 경우 문제 사이즈의 log값에 비례해서 증가함을 알 수 있다.

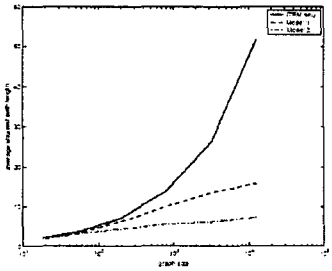

그림 2. 노드간 평균 거리의 변화

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 나와 있다. 사이즈가 커짐에 따라 복잡계 네트워크 모델을 사용할 경우 수렴속도가 크게 빨라짐을 알 수 있다.

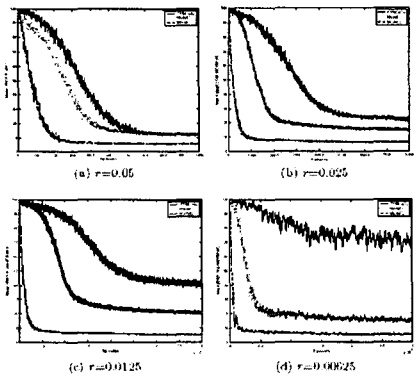

그림 3 문제 사이즈에 따른 학습 곡선

평균 학습 시간을 최종 수렴값의 절반에 다다르기까지의 세대수로 정의하였다. 그림 4는 문제 사이즈에 따른 평균 학습 시간을 나타낸다. 복잡계 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 척도 없는 도수 분포를 가짐으로써 별도의 탐색 알고리즘 없이 효율적인 탐색이 가능해지기 때문이라고 생각된다.

5. 결론

본 논문에서는 강화 학습의 문제점을 해결하기 위해 제안된 복잡계 네트워크를 사용한 환경 모델을 사용하여 강화 학습을 수행하였고, 예상대로 문제 사이즈가 커질 경우 발생하는 성능 열화가 크게 줄어듦을 알 수 있었다. 또한 평균 거리의 변화 모양이 유사한 두 모델이 실제 강화 학습시의 성능차이가 큰 것에서 효율적인 탐색 알고리즘의 중요성을 볼 수 있었다.

차후 과제로는 각 모델에 적합한 보다 우수한 탐색 알고리즘을 사용함으로써 강화 학습의 성능을 보다 끌어올리고, 이를 실제 문제에 적용하는 것을 생각할 수 있다.

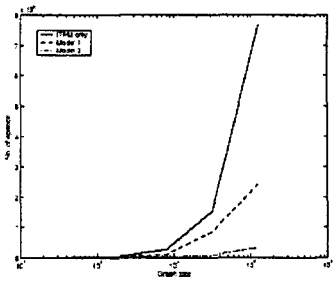

그림 4 평균 학습 시간의 변화

감사의 글

이 논문은 교육인적사업부의 BK21 사업과 산업자원부에 의해 지원되었음.

참고 문헌

[1]Barabasi, A.-L., Albert, R. Emergence of scaling in random networks. Science, 286, pp. 509-512,. 1999.

[2]Barto, A.G., Mahadevan, S. Recent advances in hierarchical reinforcement learning. Discrete Event Systems Journal, 13, 41-77, 2003.

[3]Kleinberg, J. Small-world phenomena and the dynamics of information.

[4]Klemm, K. M.guiluz, V. Growing scale-free networks with small world behavior. Phys. Rev. E, 65, 237-285, 2002.

[5]Millan, D.R., Posenato, D., Dedieu, E. Continous-action q-learning. Machine Learning, 49, 241-265, 2002.

[6]Sutton, R.S., Precup, D., Singh, S.P. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence, 112, 181-211, 1999.

[7]Watts, D.J., Strogatz, S.H. Collective dynamics of 'small-world' networks. Nature, 393, 404-407, 1998.

[8]이승준, 장병탁. 복잡계 네트워크를 이용한 강화 학습에서의 환경 표현. 한국정보과학회 봄 학술발표 논문집 (B), 제 31권 1호, pp.622-624,2004