복잡계 네트워크를 이용한 강화 학습에서의 환경 표현

이승준° 장병탁
서울대학교 바이오지능연구실
{sjlee°, btzhang}@bi.snu.ac.kr

World Representation Using Complex Network for Reinforcement Learning

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) 였다. 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.

최근 이루어진 복잡계 네트워크(Complex Network)에 대한 연구에 착안하여 본 논문은 자기조직화하는 생장 네트워크(Self organizing growing network) 를 기반으로 한 간단한 환경 표현 모델을 사용하는 강화 학습 알고리즘을 제안한다. 네트웍은 복잡계 네트웍이 갖는 성질들을 유지하도록 자기 조직화되고, 노드들 간의 거리는 작은 세상 성질(Small World Property)에 따라 전체 네트웍의 큰 사이즈에 비해 짧게 유지된다. 즉 판단해야할 단계의 수가 적게 유지되기 때문에 이 방법으로 차원성의 저주를 피할 수 있다.

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 상태의 차원이 매우 많거나 연속적인 상태를 가진 대부분의 실세계 문제에서는 모든 상태-행동의 평가값을 구하는 것이 불가능하다. 상태가 이산적인 경우에도 문제가 커짐에 따라 모든 상태-행동의 평가값을 구하는 것은 현실적으로 어려워지게 된다. 따라서 테이블의 형태가 아닌 보다 간단한 환경 표현 방식이 필요하게 된다.

일반적으로 쓰이는 해결책은 신경망과 같은 함수 근사장치를 사용하는 것이다. 강화 학습에 함수 근사장치를 사용하는 방식으로 성공적인 결과가 있어 왔다. 하지만, 어떠한 간결한(compact) 표현 방식을 사용하더라도 문제가 커짐에 따라 학습해야 할 파라미터의 수가 지수적으로 증가함이 알려져 있다. 즉, 함수 근사장치를 사용하더라도 문제가 복잡해짐에 따라 나타나는 차원성의 저주는 피할 수 없다 [2].

이 차원성의 저주는 행동의 매 단계에서 판단을 해야 하기 때문에 나타난다. 따라서 이를 피하기 위한 방법으로 한 번의 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].

이러한 중요한 성질을 이용하기 위해서, 본 논문에서는 강화 학습 문제를 네트워크에서의 정보 전달로 보는 새로운 관점을 제안한다. 강화 학습 문제의 환경인 마르코프 결정 프로세스는 각 노드가 상태를 나타내고 에지가 행동을 나타내는 네트워크로 볼 수 있으며, 보상에 대한 정보가 모든 상태로 전달되면 정책이 학습되게 된다. 네트워크를 보완하여 네트워크가 '작은 세상' 성질을 가지게 하게 하면 문제가 커지더라도 판단 단계의 수가 크게 늘어나지 않게 할 수 있으므로 차원성의 저주를 피할 수 있다. 본 논문에서는 작은 세상 성질을 유지하며 자라나는 네트웍 형태의 함수 근사 장치를 사용하는 강화 학습을 제안한다. 이러한 방식으로 사전에 문제에 대한 지식 없이 차원성의 저주를 피하면서 간결한 환경 표현이 가능하다.

2. 관련 연구

2.1. 복잡계 네트워크

복잡계 네트웍에 대한 최근 연구 결과에 의하면 대부분의 실세계 네트웍들은 다음 세 가지 성질을 가진다 [4]

작은 세상 성질 (Small world property)

네트워크 사이즈가 증가하더라도 임의의 두 노드 간에 상대적으로 짧은 경로가 존재한다는 것을 뜻한다. 대중적으로 잘 알려진 예로는 '여섯 단계의 분리(Six degrees of separation)'이다. 이것은 전 세계의 대부분의 사람들은 서로 6명 이하의 지인들로 연결이 가능하단 것이다.

높은 클러스터 계수 (High cluster coefficient)

공통된 이웃을 가지는 두 노드는 그렇지 않은 두 노드에 비해 서로 연결되어 있을 확률이 매우 높다는 것을 나타낸다. 이 성질이 무작위 네트웍과 현실 세계의 네트웍간의 큰 차이이다.

척도 없는 도수 분포 (Scale-free degree distribution)

각 노드들이 가지는 링크 수의 분포가 크기 변화에 변화받지 않는 지수적 분포를 나타낸다는 것이다. 이 성질은 네트웍의 계층적 구조와 자주 연관된다.

2.2. 복잡계 네트워크 모델

많은 실세계 네트워크들이 위의 성질들을 보이고 있기 때문에 이러한 성질을 가지는 네트워크를 모델링하려는 많은 시도가 있어 왔다.

[7]에서는 각 노드들이 이웃하고만 연결되어 있는 바둑판 모양의 격자에서 출발해서 임의의 확률로 링크를 추가한다. 결과적으로 나타나는 구조는 높은 클러스터 계수를 가지면서도 임의의 두 노드간에 높은 확률로 짧은 길이의 경로가 존재하게 된다. 하지만 이 모델은 척도 없는 도수 분포를 나타내지는 못한다.

이 도수 분포를 모델링하기 위해 [1]에서는 '부익부 빈익빈' 모델을 사용한 생장 네트워크 모델을 제안하였다. 새로운 노드가 추가될 때마다 기존의 노드들이 가지고 있는 에지의 수에 비례한 확률로 새로운 에지가 추가되게 된다. 결과적으로 척도 없는 도수 분포와 작은 세상 성질을 얻을 수 있으나, 실험적으로 얻은 모델보다 클러스터 계수가 낮다는 것이 지적되었다.

[4] 에서는 이를 보완하여 각 노드에 상태 변수를 추가함으로써 세 가지 성질을 모두 만족시키는 그래프 모델을 제안하였다.

2.3. 복잡계 네트워크에서의 탐색

네트워크에 짧은 경로가 존재한다 해도 그것을 찾을 수 없다면 의미가 없다. 하지만 사람의 경우 개인이 전체 네트워크를 모르더라도 어느 경로가 더 가능성이 있는지 판단함으로써 그 짧은 링크를 찾아낼 수 있다. 이 사실에 기인해서 복잡계 네트워크에서 효율적인 비 중앙집중적 검색 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 네트웍이 척도 없는 도수 분포를 가진다면 효율적인 검색이 가능하다는 것도 알려져 있다.

3. 복잡계 네트워크를 사용한 강화 학습

3.1. 강화 학습과 네트워크에서의 정보 전달

강화 학습 문제는 각 노드가 상태를 나타내고 에지가 행동을 나타내는 네트워크 상에서 보상에 대한 정보를 모든 상태로 전달시키는 문제로 대응이 가능하다. 즉, Q-learning과 같은 강화 학습 알고리즘은 각 행동에 대한 평가치를 주위의 상태의 값을 바탕으로 수정해 나가는 방식으로 학습을 행하는데, 이는 보상에 대한 정보를 평가치의 형태로 전파시켜 나가는 것으로 볼 수 있다.

3.2. 네트워크를 사용한 환경 표현

강화 학습에서 가정하고 있는 마르코프 결정 프로세스 자체를 네트워크의 형태로 표현 가능하다. 하지만 실제 사이즈가 큰 문제의 경우 모든 상태와 행동들을 포함해 네트워크로 만들 수 없으므로 단순한 환경 표현이 요구된다.

본 논문에서는 자기 조직화하는 성장 신경망인 ITPM[5]을 사용하여 환경을 근사하여 네트웍의 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 연결상태와 위치를 수정한다.

ITPM이 사용하는 자기조직화 알고리즘은 다음과 같다.

관측된 상태 x'에서 가장 가까운 두 노드가 b',b''라면.
1. 새로운 노드 u가 추가되었다면
   (a)u와 b',b''를 연결한다.
   (b)b'와 b''가 연결되어 있다면 에지를 삭제한다.
   추가되지 않았다면
   (c)b'와 b''를 연결한다.
2. b'와 b'의 모든 주위 노드들을 x'쪽으로 이동시킨다.

$$\Delta w_b \leftarrow \delta(x' - w_{b'})$$

$$\Delta w_r \leftarrow \delta_r(x' - w_r)$$

이 과정을 거쳐 ITPM은 각 노드가 일정 개수의 이웃과만 연결되어 있는 격자 모양 네트워크를 형성하게 된다.

3.3. 복잡계 네트워크 구현

앞서 살펴본 복잡계 네트워크의 모델들을 이 ITPM에 사용하여 환경을 복잡계 네트워크로 근사할 수 있다. 즉 ITPM은 성장 네트워크 모델이고 격자 모양 네트워크를 형성하기 때문에 앞서 설명한 모든 모델을 다 적용 가능하다. 본 논문에서는 그 중 [7]를 기반으로 한 내부 구조를 이용한 보완된 격자 모델과 [1]를 기반으로 한 모델을 적용해 보았다. 위의 자기조직화 알고리즘에 다음과 같은 부분을 추가하여 추가적인 링크를 형성한다.

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

3.4 복잡계 네트워크에서의 강화 학습

원래의 네트웍에 링크가 추가된 네트웍은 계층적 강화 학습의 보완된 MDP와 대응된다. 즉 계층적 강화 학습의 알고리즘을 사용하여 근사된 복잡계 네트워크 상에서 강화 학습을 행할 수 있다. [6] 에 따르면 각 링크 가 수행 시간 $k(s,o)$를 가진다면 각 링크들의 평가치를 다음과 같이 수정 가능하다.

$Q(s,o) \leftarrow$

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

4. 실험 및 결과

제안된 두 가지 모델과 결합된 개량된 ITPM 알고리즘을 사용하여 2차원 상의 문제에 적용하여 보았다. 에이전트는 1x1 공간에서 한번에 0.01씩 상하 좌우로 이동한다. 노드 생성시 20%의 확률로 추가적인 링크가 생성되었고, ITPM에 사용된 파라미터는 다음과 같다.

$r^2 = 0.0009, \delta = 0.0002, \delta_r = 0.00002, p = 2.322$

4.1 복잡계 네트워크 형성

정사각형의 개방된 공간에서 ITPM을 사용해서 얻은 네트워크와 제안된 두 모델을 사용시 생성된 추가 네트워크의 모양은 그림 1과 같다. ITPM만 사용시 차수가 5에 몰려 있는 균일한 격자 모양의 그래프를 얻었고 Model 1 사용시 긴 엣지들로 작은 세상 성질들이 관측되었으며, Model 2는 척도 없는 도수 분포를 보여 주었다.

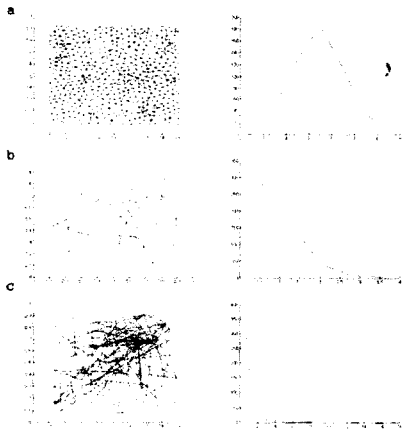

그림 1. 생성된 네트워크와 도수 분포.

4.2 노드 간 거리 비교

미로 형태의 환경 하에서 각 알고리즘을 사용하여 네트워크를 학습하였다. ITPM으로 학습된 미로의 모양은 그림 2와 같다. 세 경우에 대해 미로 내의 임의의 두 점들의 거리 분포는 그림 3과 같다. 표 1에서 볼 수 있듯이 약 10% 정도의 추가된 링크로도 평균 거리가 크게 줄어드는 결과를 얻을 수 있었다.

	ITPM	ITPM+모델1	ITPM+모델2
링크의 수	1024	1111	1128
평균 거리	14.9936	7.6451	4.4808

표 1. 링크 수와 평균 최단 거리 비교

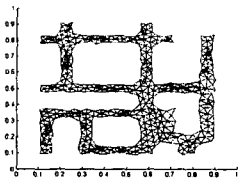

그림 2. 학습된 미로

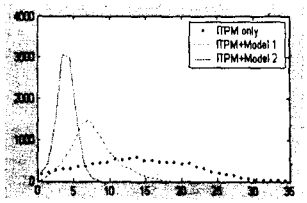

그림 3. 최단 거리 분포

5. 결론

본 논문에서는 강화 학습의 문제점 중 하나인 차원성의 저주와 이를 해결하기 위해 만들어진 계층적 강화 학습의 문제점인 간결한 환경 표현의 부재와 사전 지식이 요구되는 점을 해결하기 위해 최근 대두되고 있는 복잡계 네트워크의 성질을 응용하였다. 이를 사용해 기존 강화 학습 알고리즘의 문제점들에 대한 대안을 제시했으며, 나아가 강화 학습 문제를 네트워크에서의 정보 전파 문제로 봄으로써 양자의 방법론을 사용하여 보다 나은 결과를 기대할 수 있다.

향후 연구 방향으로는 복잡계 네트워크에서 네트워크의 구조를 사용한 적극적인 탐색 방법을 도입하는 것을 생각할 수 있다. 이러한 기법을 사용함으로써 사이즈가 큰 문제에 강화 학습을 성공적으로 적용할 수 있으리라 기대된다.

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.