탐색 강화 계층적 강화 학습

이승준[0] 장병탁
서울대학교 컴퓨터공학부 바이오지능 연구실
{sjlee,btzhang}@bi.snu.kr

Hierachical Reinforcement Learning with Exploration Bonus

Seung-Joon Yi[0] Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

Q-Learning과 같은 기본적인 강화 학습 알고리즘은 문제의 사이즈가 커짐에 따라 성능이 크게 떨어지게 된다. 그 이유들로는 목표와의 거리가 멀어지게 되어 학습이 어려워지는 문제와 비 지향적 탐색을 사용함으로써 효율적인 탐색이 어려운 문제를 들 수 있다. 이들을 해결하기 위해 목표와의 거리를 줄일 수 있는 계층적 강화 학습 모델과 여러 가지 지향적 탐색 모델이 있어 왔다. 본 논문에서는 이들을 결합하여 계층적 강화 학습 모델에 지향적 탐색을 가능하게 하는 탐색 보너스를 도입한 강화 학습 모델을 제시한다.

keywards: Q-Learning, Hierachical Reinforcement Learning, Multi-step action, Exploration bonus.

1. 서 론

강화학습은(reinforcement learning) 동적인 환경 하에서 시행착오를 거쳐 환경으로부터 주어지는 보상(reward)을 최대화하기 위한 학습 방법이다. 이러한 강화학습은 동물 행동 심리학과 최적 제어 이론 분야에 뿌리를 두고 있으며, 여러 분야에 적용되고 있다.

강화학습의 목적은 환경으로부터 주어지는 보상을 최대화하는 것이며, 이를 위해서는 Exloration과 Exploition의 적절한 조화가 필요하게 된다. 또한 실제 문제에 적용하기 위해서는 큰 상태 공간에 비해 상대적으로 적은 학습 데이터의 문제(sparse data problem)을 해결할 필요가 있다. 본 논문에서는 이러한 두 문제를 해결하기 위한 시도로써 계층적 강화 학습(Hierachical RL) 에 탐색 보너스를 도입한 탐색 강화 계층적 강화 학습 모델을 제시한다.

2. 강화 학습과 Q-Learning

Q-Learning [1]은 현존 강화 학습 방법들 중 대표적으로 쓰이는 방법으로써 시간 변화에 따른 적합도 차이를 학습에 이용하는 TD-Learning의 한 종류이다. 이 방법은 모델의 정보 없이 행동의 적합성을 나타내는 Q값만을 학습하므로 구현하기 간단하며 실제 여러 문제에 사용되어 좋은 결과를 보이고 있다.

강화 학습 환경 내에서 행동하는 에이전트(Agent)는 특정 상태에서 가능한 행동들 중 하나를 택해 행하고 다른 상태로 이동하게 된다. 이동하면서 환경으로부터 행동의 대가에 해당하는 보상(reward)를 받게 된다. 강화학습의 목표는 이러한 보상의 총합을 최대화하는 것이다. 구체적으로 매 순간 다음의 값을 최대화할 수 있도록 행동을 선택하는 것이다.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad \text{(식 1)}$$

Q-Learning에서는 아래에 정의된 optimal Q-value $Q^*(s,a)$를 직접 학습한다. 이 값은 상태 s에서 행동 a를 취한 후 최적으로 행동했을 경우의 보상의 총합을 나타낸다.

$$Q^*(s,a) = E\{ r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a') | s_t = s, a_t = a \} \quad \text{(식 2)}$$

Q-Learning의 한 step은 다음과 같이 이루어진다.

1. 현재 상태를 s라 하자.
2. 행동 a를 선택한다.
3. a를 행해서 받은 보상을 r,다음 상태를 t라 하면
4. Q(s,a)를 다음과 같이 수정한다.

$$Q(s,a) = (1-\alpha) Q(s,a) + \alpha (r + \gamma \max_{a'} Q(t,a'))$$

그림 1. Q-Learning

여기서 행동 선택은 일반적으로 ε-greedy방식이나 Boltzmann exploration을 사용한다[7]. ε-greedy방식은 일정한 확률로 임의의 행동을 선택하고 그 외에는 현재까지 알려진 가장 좋은 행동을 선택하는 것이다. (Q값이 가장 높은 행동) Boltzmann exploration은 온도 T를 사용하여 다음의 확률 $\Pr(a)$에 비례하여 행동을 선택한다.

$$\Pr(a) = \frac{e^{Q(s,a)/T}}{\sum_{a} e^{Q(s,a')/T}} \quad \text{(식 3)}$$

T가 0에 가까울 경우에는 알려진 것 중 가장 좋은 행동이 대부분 선택되게 되고 T가 클 경우에는 임의 선택에 가깝게 행동이 선택되게 된다.

모든 행동이 무한히 시행되며 학습률 α를 적절히 줄이

며 학습시킬 경우 $Q(s,a)$는 모든 s,a에 대해 optimal Q-value 인 $Q^*(s,a)$에 수렴한다는 것이 증명되어 있다[3].

3. Exploration in Q-Learning

앞서 살펴본 바와 같이 일반적인 Q-Learning은 행동을 선택하는 데에 있어서 현재의 Q값 외의 별도의 탐색 정보를 사용하지 않는다. ε-greedy 방법에서는 탐색은 완전히 무작위하게 일어나고, Boltzmann exploration에서는 현재 좋다고 알려진 행동에 치우쳐 탐색하게 된다. 즉 초기의 행동 선택에 따라 특정 방향으로 탐색이 치우칠 수가 있다. 이러한 단점을 방지하기 위해 현재 가지고 있는 Q 값 외에도 기타 정보를 사용하여 탐색에 도움을 주려는 시도가 있어 왔다.

기본적인 Q-Learning에서는 Q-value를 하나의 값으로 저장한다. Interval estimation [5], Bayesian Q-Learning [3] 등에서는 하나의 값 대신 Q-value의 분포를 사용하여 탐색에 사용한다. 즉 Q값의 불확실도가 높을 경우 낮을 경우보다 더 탐색을 함으로써 확실하지 않은 부분을 더 탐색하게 하는 방법이다. 학습이 진행됨에 따라 불확실도가 낮아지면 최적으로 보이는 행동을 행함으로써 exploitation을 행하게 된다. Interval estimation에서는 Q값의 평균이 아닌 특정 신뢰 구간의 상한을 기준으로 행동을 선택하게 된다. Bayesian Q-Learning에서는 유사한 방법으로 행동을 선택하고 더 나아가 Q값의 분포를 역전파시킴으로써 전역적 탐색을 가능하게 하지만 계산 부담이 매우 큰 단점이 있다.

다른 방법으로는 명시적인 탐색 보너스(exploration bonus)를 사용하는 방법이 있다[6]. 이 방법에서는 상태의 방문 회수나 예측 오차 등의 값을 별도로 저장한 후 최근에 덜 방문된 상태나 예측 오차가 큰 상태에 탐색을 집중시키는 방법이다. 비교적 간단하게 구현 가능한 장점이 있다. 예로 다음과 같은 탐색 보너스가 사용 가능하다.

$$E^{counter}_{explore}(a) = \frac{1 + c(s_t)}{1 + c(\hat{s_{t+1}}(s_t, a))} \quad \text{(식 4)}$$

4 Hierachical Q-Learning

Q-Learning과 같은 기본적인 강화 학습 방법은 문제의 사이즈가 커짐에 따라 성능이 매우 떨어지게 된다. 한 가지 이유는 문제가 커짐에 따라 행동과 보상간의 거리가 늘어나는 데에 있다. 보상이 다시 원래 행동이 행해진 상태에까지 역전파되어 와야 하기 때문이다.

이러한 문제를 해결하기 위해 상태의 계층화[4]가 사용된다. 이러한 상태의 계층화에서는 보다 높은 단계에서 하위의 많은 단계를 건너뛰는 것이 가능하므로 행동과 보상간의 거리를 문제가 커지더라도 대응이 가능하다. 기존의 상태 계층화 방법은 다음과 같은 것들이 있다.

4-1. 정의된 상위 행동(abstract action)을 사용하는 경우

이 경우는 행동들의 묶음으로 이루어진 상위 행동이 미리 정의되어 있을 경우이다. 이 경우 상위 행동을 사용하게 되면 상태 공간이 크게 줄어들게 되어 행동에서 보상까지의 거리가 크게 짧아지게 된다. 즉 학습이 현저하게 쉬워지게 된다[7].

4.2 정의된 하위 목표(subgoal)를 사용하는 경우

구체적인 행동의 위계 관계를 제시하는 대신 최종적인 목표를 여러 개의 하위 목표(subgoal)로 분해해 제시해 주는 방법이다. 이 경우 상위 행동은 각 하위 목표들을 별도로 푸는 방법으로 학습된다. 역시 작은 크기의 하위 목표들을 푼 뒤 종합하는 방법을 사용하면 전체 문제의 학습이 보다 간단해지게 된다[2].

4.3 다단계 행동(Multi-step action)을 사용하는 경우

위의 두 가지가 모두 정의되지 않을 경우에는 4-1의 변형으로 다단계 행동을 사용하는 강화 학습이 가능하다. 다단계 행동은 행동들을 여러 개 묶어서 하나의 행동 단위로 사용하는 것이다. 정의된 상위 행동을 사용하는 경우보다는 임의의 다단계 행동을 상위 행동으로 사용한다는 점에서 비효율적일 수 있으나 보다 구현이 간단하고 임의의 문제에 적용이 가능하고 여러 가지 스케일의 행동들을 동시에 사용한 학습이 가능하다는 장점이 있다[4].

5. Hierachical Q-Learning with Exploration Bonus

본 논문에서는 앞서 말한 탐색과 scaling의 문제를 해결하기 위해 계층적 강화 학습 모델에 탐색 보너스를 적용한 탐색 강화 계층적 강화 학습 모델을 제시한다. 일반적인 문제에 적용 가능하도록 하고 간단한 구현이 가능하도록 하기 위하여 다단계 행동(Multi-step action)을 사용한 계층적 강화 학습 모델에 상태 카운트를 사용한 전역적 탐색 보너스를 사용하였다. 이 모델의 경우 행동 선택은 다음과 같이 탐색 보너스를 사용하여 이루어진다.

$$\begin{aligned} a_{select}(s_t, a) &= argmax_a(Q_{combined}(s_t, a)) \\ &= argmax_a(Q'(s_t, a) + \beta E^{counter}_{explore}(\hat{s_{next\_t}}(s_t, a))) \\ &= argmax_a(Q'(s_t, a) + \beta \frac{1 + c(s_t)}{1 + c(\hat{s_{next\_t}}(s_t, a))}) \end{aligned} \quad \text{(식 5)}$$

Multi-Step Action $a = a_1, a_2, \ldots, a_n$ 을 행했을 경우는

$$Q'(s_t, a) = \frac{1}{\Sigma_{k=1}^{n} \gamma^{k-1}} \Sigma_{k=1}^{n} \gamma^{k-1} Q(\hat{s}_{t+k-1,} a_k) \quad \text{(식 6)}$$

여기서 β는 Q값과 탐색 보너스 간의 균형을 위한 상수이고 next_t는 a를 취한 후의 time step이다. MSA를 취할 경우 next_t =t+n이 된다.

행동을 취한 후의 Q-value update는 일반 Q-Learning과 동일하게 이루어진다. 즉 다음 식에 의해 이루어진다.

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(t,a')) \quad \text{(식 7)}$$

Multi-step action을 취했을 경우 역시 각 단위 행동을 n번 행한 것과 같이 Q-value가 update되지만 순서가 역순으로 행해진다. 즉 나중에 방문된 state처럼 Q-value가 update됨으로써 Q값의 역전파가 빠르게 일어날 수 있도록 된다. 도중에 지난 상태 s에 대해서는 카운터 c가 1씩 증가하게 된다.

6. 실험 및 결과

실험은 다음과 같은 11*11 Gridworld 내에서의 탐색 문제로 행해졌다. 보상은 Goal state에서 1, 기타 모든 state에서 0이다. 즉 이 강화학습 문제에서 보상을 최대화하는 정책은 Start에서 Goal까지의 최단 경로가 된다.

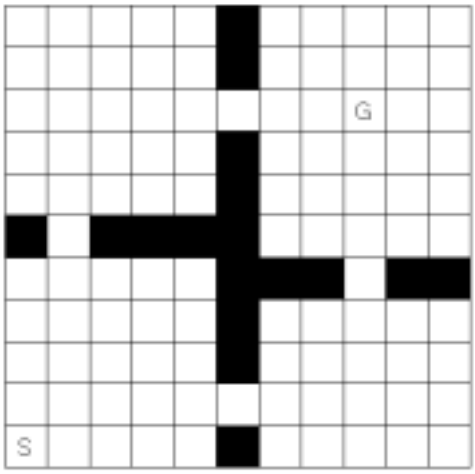

그림 2. Gridworld

6-1. Random Walk Case

이 실험에서는 MSA와 Exploration bonus, 그리고 이 둘을 복합적으로 사용한 경우의 탐색 능력을 알아보았다. 시작점에서 고울 지점까지 어느 정도의 시간이 흐른 다음에 도착하는지의 평균값 (First Passage Time, FPT)를 측정하여 탐색 능력의 기준으로 삼았다. 비교 대상은 Random walk, Random walk with MSA, Random walk with Exploration bonus, Random walk with MSA and exploration bonus 의 네 가지였다. 1000번 시행해 구한 결과는 다음의 표와 같다. (MSA는 2-step 까지 사용하였다)

탐색방법	Random walk	RW w/MSA	RW w/E.bonus	RW w/both
FPT	780	613	207	195

표 2 Random-walk의 결과

Random walk로 Start 지점부터 Goal까지 가는 데에는 평균 780회의 행동이 필요하였다. 2-step action을 추가한 경우 단위 행동 대비 20% 정도의 향상을 보였다. 2-step action을 단일 행동으로 본다면 그 이상의 향상이라고 여길 수 있다. 탐색 보너스를 사용한 방법은 완전히 Random한 방법에 비해 30% 정도의 행동으로 골에 다다를 수 있었다. 한 번 지나온 곳보다 안 가 본 곳에 탐색이 집중되기 때문에 큰 성능 차이가 있다고 생각된다. Bonus와 MSA를 동시에 사용한 경우 그에 추가하여 약간의 성능 향상을 얻을 수 있었다.

6-2 Q-Learning Case

이번에는 Random walk 대신 Q-Learning을 사용하였을 경우의 수렴 형태를 알아보았다. Q-Learning의 경우는 ε=0.5부터 0까지 linear하게 감소하는 ε-greedy 탐색을 사용하였고 γ는 0.9을 사용하였다. 1000회 반복 수행시 모두 최적치(16 step)을 찾을 수 있었다. MSA를 사용한 경우와 MSA, 보너스를 사용한 경우에는 two-step action을 최대 회수(7회) 사용하는 올바른 최적치를 구할 수 있었다. 다음은 학습 도중의 경로 길이이다.

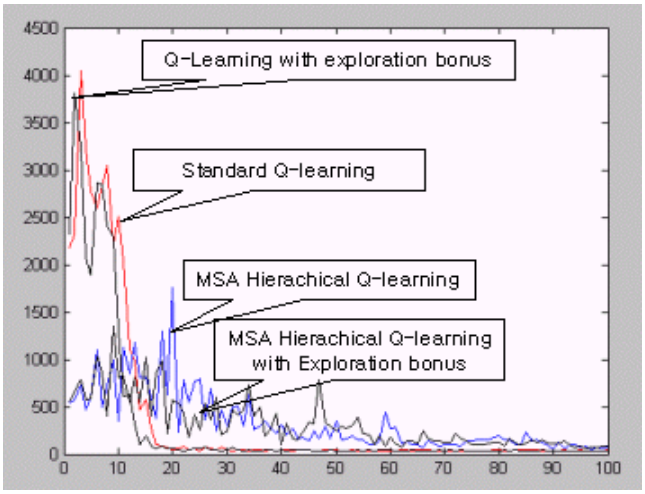

그림 3 초기 100 step 동안의 수렴

문제 size가 작아 빠른 속도로 수렴이 일어났다. Exploration bonus를 사용한 경우 사용하지 않은 경우보다 항상 더 우수한 초기 수렴 성능을 나타냈다. MSA의 경우에도 단독 사용시보다 성능 향상을 볼 수 있었다. 많은 탐색이 필요한 큰 도메인의 경우 보다 큰 성능 향상이 기대된다.

감사의 글 : 본 연구는 BK21-IT 프로젝트에 의해 일부 지원 받았습니다.

참고 문헌

[1] Watkins,C.J. and Dayan,P. Q-Learning. *Machine Learning*, 8(3):279-292, 1992.

[2] Dietterich,T.G. Hierachical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*. 1999.

[3] Dearden,R., Friedman,N. and Russell,S. Bayesian Q-Learning. Procedings AAAI-98, Madison, Wisconsin: AAAI Press, 1998.

[4] Schoknecht,R. Hierachical Reinforcement Learning with Multi-step actions. 2001.

[5] Kaelbling,L.P. *Learning in Embedded Systems.* PhD thesis, Department of Computer Science, Stanford University, 1990.

[6] Thrun,S.B. The role of exploration in learning control. In D.A. White and D.A.Sofge,eds., *Handbook of Intelligent Control:Neural,Fuzzy and Adaptive Approaches.* Van Nostrand Reinhold, 1992.

[7] Parr,R.E. *Hierachical control and learning for markov decision processes.* Doctoral dissertation, University of California, Berkley, CA. 1998.