사전 지식에 의한 강화학습 에이전트의 학습 속도와 경향성 변화

김지수*, 이은헌*, 김현철*
*고려대학교 컴퓨터학과
gameboyjisoo@korea.ac.kr, booksky@korea.ac.kr, harrykim@korea.ac.kr

How the Learning Speed and Tendency of Reinforcement Learning Agents Change with Prior Knowledge

Jisoo Kim*, Eun Hun Lee*, Hyeoncheol Kim*
*Dept. of Computer Science and Engineering, Korea University

요 약

학습 속도가 느린 강화학습을 범용적으로 활용할 수 있도록 연구가 활발하게 이루어지고 있다. 사전 지식을 제공해서 학습 속도를 높일 수 있지만, 잘못된 사전 지식을 제공했을 위험이 존재한다. 본 연구는 불확실하거나 잘못된 사전 지식이 학습에 어떤 영향을 미치는지 살펴본다. OpenAI Gym 라이브러리를 이용해서 만든 Gamble 환경, Cliff 환경, 그리고 Maze 환경에서 실험을 진행했다. 그 결과 사전 지식을 통해 에이전트의 행동에 경향성을 부여할 수 있다는 것을 확인했다. 또한, 경로 탐색에 있어서 잘못된 사전 지식이 얼마나 학습을 방해하는지 알아보았다.

키워드: 기계학습, 강화학습, 사전 지식, Q-learning, 강화 학습 에이전트

1. 서론

기계학습의 분야 중 하나인 강화학습은 trial-and-error 방식으로 주어진 환경(environment)에서 보상에 도달하는 정책(policy)을 배우는 알고리즘이다. 강화학습에서는 행동의 주체가 되는 에이전트(agent)가 가능한 행동(action) 중 한 개를 계속 선택해서 최종적으로 목표에 도달하는 것을 목표로 삼는다[1]. 단계적으로 문제를 푸는 이 방식은 현존하는 방식 중 사람의 문제풀이 방식과 가장 유사하다는 특징이 있다[2].

그러나 강화학습은 복잡한 환경에서 많은 학습시간이 필요하다는 단점을 가지고 있다[3]. 이를 해결하는 방법 중 하나가 에이전트에게 사전 지식(prior knowledge)을 알려주는 것이다[4]. Q-learning[5]을 쓰는 로봇에게 사전 지식을 알려주었을 때가 백지상태(tabula rasa)에서 학습하는 것보다 더 빠르게 학습하고 좋은 성과를 나타낸 사례가 있다[6][7]. 또한, 사람이 보여주는 선례를 모방하여 학습하는 에이전트도 좋은 결과를 보여준 적이 있다[8]. 그러나 이런 방법이 성공할 수 있었던 이유는 해당 문제에 대해 정답에 해당하는 지식을 제공할 수 있었기 때문이다[9]. 정답이 명확하지 않은 문제에 대해서는 인공지능에게 잘못된 지식을 전달할 수 있으며, 그것이 잘못된 행동으로 이어질 수도 있다[10][11]. 그렇기 때문에 이런 잠재적 위험이 어떤 경우에 발생할 수 있는지 살펴볼 필요가 있다.

이에 본 연구는 Q-learning 을 쓰는 3 가지 환경에서 이런 사전 지식의 영향력을 확인한다. 2 개의 경로 탐색 환경에서는 탐색을 방해하는 사전 지식과 차선책을 지향하는 사전 지식이 최종 경로 결정에 어떤 영향을 미치는지 본다. 도박 환경에서는 안정성 혹은 도박성을 추구하는 사전 지식의 강도에 따라서 어떻게 행동이 바뀌는지 확인한다.

본 연구의 2 장에는 관련 연구에 대해서 서술했다. 3 장에는 구축한 환경에 대한 세부 정보를 기술했으며, 4 장에는 각 환경에서 진행한 실험에 대해 설명하고 실험 결과를 분석했다. 결론에는 본 연구를 통해 얻은 결과 및 연구의 한계에 대해서 서술했다.

2. 관련 연구와 비교

Q-learning 에 사전 지식은 실험 시작 전부터 알고 있는 지식이며, 이런 사전 지식을 적용해서 학습 속도와 성능을 높인 연구가 있다. 그 중 Dixon[6]은 부분적인 사전 지식이 학습을 얼마나 돕는지 확인했다. 벽을 끼고 이동하는 로봇이 가장 흔한 12 개의 상태에 대한 정답을 알면 얼마나 더 빠르게 학습하는지

봤다. 그 결과 일부 상황에만 적용한 사전 지식이 학습 속도를 7.5 배까지 높일 수 있다는 것을 확인했다.

Moreno[7]는 복잡한 환경을 단계별로 나눠서, 술래잡기 환경에서 플레이어가 술래를 피해서 목표 지점에 도달하도록 학습시켰다. 처음에는 목표 지점만 존재하는 환경에서 학습시켰으며, 환경에 익숙해지면 술래와 플레이어 한 명을 차례대로 추가해서 학습시켰다. 그 결과 사전 지식을 쓴 플레이어의 승률이 약 2 배 높였으며, 학습 시간 또한 절반 이하로 줄였다.

두 연구를 통해서 사전 지식으로 정답을 줬을 때 학습 시간을 줄이고 성능을 높일 수 있다는 것을 확인했다. Dixon 은 사람이 제공하는 사전 지식의 효과를 확인했으며, 같은 문제를 다른 환경에서 풀 때도 범용적인 사전 지식이 쓰일 수 있다는 것을 알게 되었다. Moreno 는 일부 상황에 대한 정답을 배우면, 그것을 기반으로 더 복잡한 상황을 빠르게 학습할 수 있다는 것을 확인했다. 강화학습 에이전트도 단계별로 문제를 나눠서 풀 수 있다는 것을 검증했다.

그러나 Dixon, Moreno 모두 사전 지식으로 정답을 주는 경우에 대해서만 확인했다. 때로는 사람이 알고 있는 사전 지식이 잘못됐을 수도 있으며, 학습에 도움이 안되는 문제의 일부를 먼저 학습시킬 수 있다. 또한, 정답이 없는 문제에 대해서 확인하지 않았다. 그렇기에 본 연구는 잘못되거나 경향성을 주는 사전 지식이 어떻게 학습에 영향을 주는지 확인한다.

3. 환경 소개

오픈소스 라이브러리인 OpenAI Gym[12]의 Kelly Coinflip, Cliff, Frozen Lake 환경을 참고해서 최종적으로 Gamble, Cliff, Maze 환경을 만들었다. 모든 환경에서 처음에는 탐험을 우선시할 수 있도록 무작위성을 어느 정도 부여하고, 학습이 진행되면서 무작위 행동을 취할 확률을 점차 감소시켰다. 또한 모든 환경에서 할인계수(discount rate)는 0.97, 학습률(learning rate)은 0.01 로 고정했다.

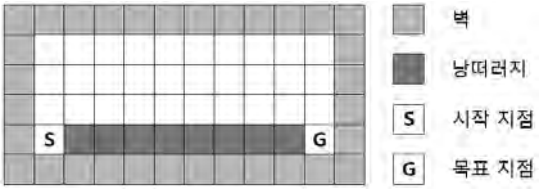
경로 탐색 환경에서는 목표 지점까지 가는 최적 경로를 찾을 수 있도록 이동에 대한 부정적인 보상을 적용했다. 한 번 행동을 취할 때 보상을 -1 로 지정함으로 에이전트가 효율적으로 이동하도록 유도했다.

3.1 Gamble 환경

Gamble 환경에서는 동전 뒤집기로 도박하면서 최대한 돈을 모아야 한다. 보유한 자산의 일정 비율을 베팅하고, 앞면이 나오면 베팅 금액만큼 얻고 뒷면이 나오면 베팅 금액만큼 잃는다. 한 학습 회차는 총 10 라운드 동안 진행되며, 그 전에 돈을 다 잃거나 최대 금액까지 벌면 회차가 종료된다. 초기 자금은 20 원, 동전 앞면 확률은 0.62, 최대 금액은 150 원으로 책정했다. 그 이유는 해당 초기값으로 했을 때, 에이전트의 성향이 조금씩 다르게 나왔으며, 사전 지식으로 행동이 어떻게 변하는지 알아볼 수 있었기 때문이다.

현재 보유 금액을 5 로 나눈 것(나머지는 버린다)이 에이전트의 상태가 되고, 보유 금액에 따라서 총 30 개의 상태가 존재한다. 가능한 행동은 보유 자산의 10%, 20%, 40%, 60%를 도박하는 것으로 총 4 가지가 있다. 처음에 무작위 행동을 취할 확률을 100%로 설정했으며, decay rate 를 0.999 로 설정했다.

3.2 Cliff 환경

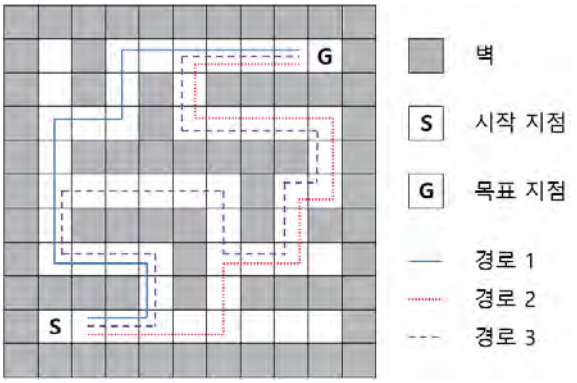


(그림 1) Cliff 환경

낭떠러지를 피해서 목표 지점으로 가는 경로 탐색 환경이다(그림 1). 목표 지점에 도달하면 +30 보상을 얻고, 낭떠러지에서 떨어지면 -30 보상을 얻으면서 학습 회차가 종료된다. 제자리걸음을 방지하기 위해서 벽에 부딪히면 -10 의 보상을 얻게 했다.

에이전트의 위치가 현재의 상태이며, 총 72 개의 상태가 존재한다. 그 중 28 개는 벽이며 실제로 사용하는 상태는 44 개다. 가능한 행동은 상하좌우 이동 네 가지다. 처음에 무작위 행동을 취할 확률을 60%로 설정했으며, decay rate 를 0.99 로 설정했다.

3.3 Maze 환경



(그림 2) Maze 환경

Cliff 환경과 같이 경로 탐색 환경이다 (그림 2). 목표 지점에 도달하면 +30 보상을 얻고, 제자리걸음을 방지하기 위해서 벽에 부딪히면 -20 의 보상을 얻는다.

경로 1 은 최적 경로로 22 번의 행동 후에 목표 지점에 도달하며, 경로 2, 3 은 각각 24 번, 34 번 만에 목표 지점에 도달한다.

에이전트의 위치가 현재의 상태이며, 총 121 개의 상태가 존재한다. 그 중 70 개는 벽이며 실제로 사용하는 상태는 51 개다. 가능한 행동은 상하좌우 이동 네 가지다. 처음에 무작위 행동을 취할 확률을 10%로 설정했으며, decay rate 를 0.99 로 설정했다.

4. 실험 및 실험 결과

사전 지식이 에이전트에 미치는 영향을 보기 위해서 최적화를 돕는 사전 지식, 최적화를 방해하는 사전 지식, 차선책으로 유도하는 사전 지식, 그리고 경향성을 부여하는 사전 지식을 점차 적용해봤다.

<표 1> Gamble 환경에서의 각 베팅 비율에 대한 사전 지식 적용 결과

	사전 지식 無		10% 베팅 유도		20% 베팅 유도		40% 베팅 유도		60% 베팅 유도	
사전 지식	평균	표준 편차	평균	표준 편차	평균	표준 편차	평균	표준 편차	평균	표준 편차
0.2	389.7	121.27	327.65	113.35	400.65	149.68	374	129.54	355.3	105.70
0.4			306.75	98.12	355.15	133.68	372.45	152.95	434.75	149.05
0.6			286.55	97.50	342.25	128.98	423.75	141.79	457.9	148.20
0.8			277.45	107.23	351.85	100.52	358.5	123.64	440.25	211.78
1.0			265.95	66.13	320.55	91.16	426.35	139.19	455.75	172.02

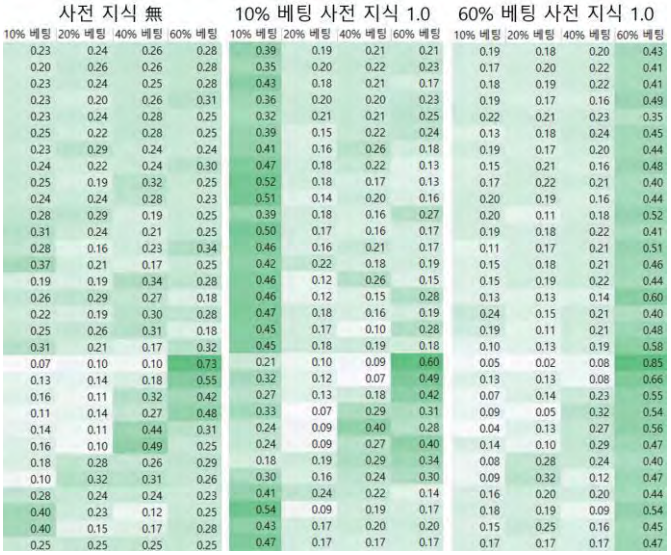
원래 모든 값이 0 으로 시작하는 Q-table 의 일정 구간에 양수값, 음수값을 적용하는 방식으로 실험에서 사전 지식을 부여했다.

4.1 Gamble 환경

총 300 번 학습시킨 후 생성된 Q-table 을 기반으로 게임을 플레이하도록 했다. 모든 테스트는 플레이 20 회로 진행됐으며, 평균과 표준편차를 비교했다.

사전 지식이 없을 때, Q 값이 약 0.2~5 사이에 머물렀다. 또한 평균 389.7 원을 모은 상태로 게임이 종료되었으며, 게임 간의 표준편차는 약 121.27 이었다.

사전 지식의 정도에 따라서 에이전트의 행동이 어떻게 바뀌는지 살펴보았다. 돌아가면서 한 행동에 대해 0.2, 0.4, 0.6, 0.8, 1.0 의 Q 값을 적용했다. 경향성의 변화를 나타낸 결과는 표 1 과 그림 3 에 나타냈으며, 소수점 아래 2 자리까지 반올림했다.



(그림 3) Gamble 환경에서 사전 지식으로 행동 선택에 변화를 대표적으로 나타낸 사례들의 히트맵

사전 지식을 강하게 부여하면 에이전트의 행동에 경향성을 부여할 수 있다는 것이 확인되었다. 특히 안전성을 추구하는 10% 베팅과 도박성을 추구하는 60% 베팅에 대한 사전 지식은 확실한 변화를 보여준다. 사전 지식을 강하게 적용했을 때는 거의 항상 해당 행동을 선택하는 것이 확인되었다.

안전성을 추구하는 사전 지식이 강하면 평균 총수익이 약 30%까지 줄지만, 표준편차는 절반까지 감소한다. 도박성을 추구하는 사전 지식이 강할수록 평균 총수익이 약 17% 늘지만, 표준편차도 2 배까지 는다.

4.2 Cliff 환경

처음으로 최적 경로를 찾는 학습 회차를 확인하는 실험을 진행했다. 모든 테스트는 플레이 20 회로 진행됐으며, 평균과 표준편차를 비교했다.

사전 지식이 없을 때, Q 값이 약 -9 에서 20 사이에 머물렀다. 목표 지점에 가까울수록 양수 값이 컸으며, 낭떠러지 직전에서 낭떠러지로 이동하는 행동에 대한 Q 값이 가장 낮았다. 평균적으로 300 번째 회차에 학습이 됐으며, 표준편차는 약 92.53 이었다.

긍정적, 부정적 사전 지식의 정도에 따라서 학습 속도가 얼마나 바뀌는지 살펴보았다. 낭떠러지로 떨어질 수 있는 8 개의 상태에 최적 경로를 가르치는 사전 지식(우측으로 이동)과 낭떠러지로 유도하는 사전 지식(아래로 이동)을 적용했다. 또한, 이 두 가지 중 한 가지에 양수의 Q 값을 적용하면, 다른 경우에 대해서 같은 Q 값을 음수로 적용했다.

여러 테스트를 해본 결과, 0.1~0.6 사이의 Q 값들이 가장 유의미한 결과를 나타냈기 때문에 해당 값으로 최종 실험을 진행했다. 최대 1000 번까지 학습을 할 수 있도록 설정했으며, 그때까지 학습이 안 되면 실패했다고 간주했다. 결과는 표 2 에 나타냈으며, 소수점 아래 2 자리까지 반올림했다.

<표 2> 최적 경로와 낭떠러지로 떨어지는 경로에 대한 사전 지식을 적용했을 때 학습 결과

	최적 경로		낭떠러지		
사전 지식	평균	표준 편차	평균	표준 편차	실패 횟수
0.1	148	155.18	339.1	92.92	0
0.2	46.25	115.57	339.15	78.83	0
0.3	21.7	54.14	364.2	83.78	0
0.4	9.85	6.51	358.05	77.15	0
0.5	9.8	4.88	406.38	84.06	7
0.6	8.05	4.70	439.44	111.86	11

최적 경로를 찾는데 도움이 되는 사전 지식의 경우, 적은 양을 부여하더라도 학습 속도를 크게 높일 수 있었다. 낭떠러지로 유도하는 사전 지식의 경우에도 적은 양으로도 학습 속도를 10%~20% 늦출 수 있었다. 또한, 잘못된 사전 지식이 너무 커지면 절반 이상의 경우에 학습 자체를 실패한다. 이를 통해 간단한 경로 탐색 환경에서 작은 양의 사전 지식도 큰 영향을 끼칠 수 있다는 것을 확인했다.

4.3 Maze 환경

3 개의 경로 중 한 경로로 학습을 유도하는 사전 지식을 부여했을 때 어떤 경로를 선택하는지 확인하는 실험을 진행했다. 학습과 테스트를 병행했으며, 총 3000 번 학습을 시키면서 한 경로만 계속 선택하게 되는 시점을 기록했다. 경로를 찾을 때 100 번 이상 행동을 취해도 목표 지점을 찾지 못하면, 다음 학습 회차로 넘어가도록 했다.

사전 지식 없이 5 회 테스트했을 때, 약 1780 회 학습을 진행하면 최적 경로를 확정적으로 찾기 시작했다. 이때 Q 값이 약 -10 에서 20 사이에 머물렀다. 양수 값은 목표 지점 근처라는 것을 고려하면, 중간 경로의 Q 값은 -10 까지 간다고 판단했다.

Q-learning 특성상 한 방향으로 이동하게 하기 위해서 나머지 세 방향에 음수인 Q 값을 적용했다. 그렇지 않으면, 사전 지식으로 부여한 양수의 Q 값이 나머지 3 방향과 같아질 때까지 반복적으로 사전 지식이 부여된 상태 내에서만 이동했기 때문이다.

경로 1, 2, 3 으로 유도하는 사전 지식을 적용하고 그 결과를 살펴보았다. 여러 테스트를 해본 결과, 1~30 사이의 Q 값들이 가장 유의미한 결과를 나타냈기 때문에 해당 값으로 최종 실험을 진행했다. 이때, 환경의 복잡도를 반영하여 부분적으로만 사전 지식을 적용했다. 경로 전체에 대해 사전 지식을 적용하지 않고 시작 지점부터 전체 경로의 약 3 분의 1 에 대해 사전 지식을 부여했다. 결과는 표 3 에 나타냈다.

<표 3> 경로별로 사전 지식을 적용했을 때 학습을 마치는 학습 회차

사전 지식 無	사전 지식 有	경로 1	경로 2	경로 3
1783	1	1987	1818*	1846*
1791	5	2181	1605*	1733*
1777	10	1973	1985*	1532*
1780	20	1405	1593	2387
1799	30	1414	1570	2401

* 해당 경우에는 경로 1 로 학습이 마무리되었다.

극단적이지 않은 사전 지식도 학습에 도움을 준다는 보장이 없었다. 최적 경로인 경로 1 에 대해서 -10 이상의 사전 지식을 부여했을 때만 학습 속도가 빨라지는 것이 확인되었으며, 그보다 적은 사전 지식은 학습 속도를 높이지 못했다. 나머지 경로의 경우에도 Q-table 에 일반적으로 나타나는 값 이상으로 부여했을 때만 에이전트가 선택하는 최종 경로를 바꿀 수 있었다. 또한, 경로 3 의 경우에는 극단적인 사전 지식을 적용하더라도 더 많은 학습 회차가 필요했다.

이런 결과를 통해서 Q-learning 에서 사전 지식의 한계를 살펴볼 수 있었다. 경로가 길어질수록 목표 지점의 보상이 경로 전체에 적용될 때까지 많은 학습 회차가 필요하다. 또한, 사전 지식을 부여하는 방식 때문에 적은 사전 지식이 오히려 학습을 방해하는 효과를 가져왔다.

5. 결론

본 연구는 3 개의 환경에서 진행한 Q-learning 실험을 통해서 최적화되지 않은 사전 지식의 영향에 대해 살펴보았다. 그 결과 정답이 없는 환경에서 에이전트의 행동 경향을 성공적으로 바꿨다. 또한, 부여하는 사전 지식의 정도에 따라서 행동 경향을 더 많이 바꿀 수 있다는 것도 확인했다. 경로 탐색 환경의 경우, 사전 지식이 최적 경로 탐색을 방해하면, 학습이 느려지거나 실패할 수 있다는 것을 확인했다. 그리고 환경이 복잡해지면 극단적인 사전 지식을 부여해야지만 학습에 영향을 끼칠 수 있다는 것을 확인했다.

향후에는 Q-learning 외 다른 강화학습 기법에서 정답이 아닌 여러 종류의 사전 지식이 어떤 영향을 끼치는지 살펴볼 필요가 있다. 또한, 기존 연구에 쓰였던 주변 사물을 인식하는 관계적 상태 공간에서도 다른 종류의 사전 지식의 영향을 확인할 필요하다.

참고문헌

[1] Kaelbling, L. P., Littman, M. L., & Moore, A. W. Reinforcement learning: A survey. Journal of artificial intelligence research, 4, 237-285. 1996.

[2] Sutton, R. S. and Barto, A. G. Reinforcement Learning: An Introduction. MIT Press. 1998.

[3] Driessens, K., & Džeroski, S. Integrating guidance into relational reinforcement learning. Machine Learning, 57(3), 271-304. 2004.

[4] Smart, W. D., & Kaelbling, L. P. Practical reinforcement learning in continuous spaces. ICML. 2000. 903-910.

[5] Watkins, C. J., & Dayan, P. Q-learning. Machine learning, 8(3-4), 279-292. 1992.

[6] Dixon, K., Malak, R. J., & Khosla, P. K. Incorporating prior knowledge and previously learned information into reinforcement learning agents. Carnegie Mellon University, Institute for Complex Engineered Systems. 2000.

[7] Moreno, D. L., Regueiro, C. V., Iglesias, R., & Barro, S. Using prior knowledge to improve reinforcement learning in mobile robotics. Proc. Towards Autonomous Robotics Systems. Univ. of Essex, UK. 2004.

[8] Abbeel, Pieter, and Andrew Y. Ng. Exploration and apprenticeship learning in reinforcement learning. Proceedings of the 22nd international conference on Machine learning. 2005.

[9] Argall, B. D., Chernova, S., Veloso, M., & Browning, B. A survey of robot learning from demonstration. Robotics and autonomous systems, 57(5), 469-483. 2009.

[10] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565. 2016.

[11] Everitt, T., Krakovna, V., Orseau, L., Hutter, M., & Legg, S. Reinforcement learning with a corrupted reward channel. arXiv preprint arXiv:1705.08417. 2017.

[12] https://gym.openai.com/