그리드 환경에서의 강화학습 감쇠율 변화에 따른 경로의 다양성 분석

박정우, 최재영*
가천대학교
jychoi19@gachon.ac.kr

Analysis of Diversity of Paths according to Discount Rate of Reinforcement Learning in Grid Environment

Park Jung woo, Jaeyoung Choi*
Gachon Univ.

요 약

본 논문은 강화학습에서 핵심 변수인 감쇠율(discount rate, γ)의 설정이 에이전트의 정책 학습 과정과 최종 행동 전략에 어떤 영향을 미치는지를 심층적으로 분석하였다. 기존의 강화학습 연구에서는 감쇠율을 단순히 장기 보상에 대한 가중치를 조절하는 하나의 값으로 알았으나 본 연구는 감쇠율이 단기적 또는 장기적 행동 전략을 유도하는 정책의 구조 자체에 중대한 영향을 미칠 수 있다는 전제하에 실험을 설계하였다. 실험 결과 감쇠율이 정책의 시간적 구조와 위험 회피 성향에 직접적으로 영향을 미치는 설계 변수임을 실험 결과를 통해 확인할 수 있었다. 결과적으로 본 논문은 강화학습에서 감쇠율이 단순한 학습 조절 파라미터를 넘어, 정책의 특성을 설계하는 핵심 요소라는 점을 강조하며, 실제 응용 환경에서는 목표의 성격에 따라 적절한 감쇠율 설정이 필요하다는 것을 보여준다.

Ⅰ. 서 론

강화학습은 에이전트(agent)가 환경(environment)과 상호작용하며 최적의 정책(policy)으로 행동 전략을 학습하는 구조로, 다양한 순차적 의사결정 문제에 기초를 두고 있다[3]. 특히, 현재의 중요도를 통해 미래 보상에 대한 정책의 방향성과 행동(action) 방향에 직접적으로 영향을 미치는 감쇠율이 있다[3][5]. 그러나 감쇠율은 실제 정책에서 어떤 행동 양식을 유도하는지에 대한 실증적인 분석이 충분히 이루어지지 않았다. 또한 감쇠율은 강화학습에서 미래 보상의 현재 가치를 결정짓는 핵심 매개변수로 에이전트의 행동 전략에 지대한 영향을 미친다. 일반적으로 감쇠율이 높을수록 장기적인 보상에 대한 가중치가 커지고 낮을수록 단기적인 보상에 집중하게 된다[3][5]. 그러나 감쇠율은 환경의 특성과 목표에 따라 최적값이 달라지며 확실하지 않은 설정값은 학습 속도를 저하하거나 불안정한 정책을 만들 수 있다.

이에 본 연구에서는 실내 경로 계획 문제를 예시로 설정하고, 감쇠율의 변화에 따라 학습된 정책이 어떻게 변화하는지 모의실험을 통해 분석하고자 한다. 실험을 통해 감쇠율을 주어진 환경에서 다양한 값을 설정하고 적용하여 성능 향상에 이바지할 수 있는 최적의 감쇠율을 만들고자 한다. 또한 정책의 보상 최적화와 충돌 회피 능력, 경로 다양성 등 주요 성능 지표가 감쇠율의 변화에 따라 어떻게 달라지는지를 비교 및 분석하였다.

Ⅱ. 본론

A. Q-learning 알고리즘 개요

본 논문에서는 Model free 강화학습의 대표적인 학습방법인 Q-학습을 적용하였다. 강화학습은 에이전트가 상태(state)에서 행동을 선택하고, 그에 따른 보상(reward)과 다음 상태를 관찰하면서 정책을 학습하는 과정이다. 감쇠율은 0부터 1 사이의 값을 가지며, 미래 보상을 현재의 의사결정에 얼마나 반영할지를 결정한다. 감쇠율이 낮으면 즉각적인 보상에 집중하며, 감쇠율이 높을수록 장기적인 보상까지 고려하게 된다[3][5]. 이때 Q-learning은 다음의 벨만 방정식을 기반으로 Q 값을 반복적으로 갱신하며 정책을 최적화한다[2]:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot [r + \gamma \cdot \max_{a'} Q(s',a') - Q(s,a)]$$

수식 1. r: **보상**, $\max_{a'} Q(s',a')$: **미래 최대 가치**, s': **다음 상태**, a': **다음 상태에서의 선택 가능한 행동들**

여기서 α은 학습률(learning rate), ε은 탐험 비율(exploration rate)이다. 학습률은 새로운 정보를 기존 Q 값에 얼마나 반영할지를 나타낸다[3]. 탐험 비율 ε은 무작위 행동을 시도할 확률로 ε-greedy에서는 $1-\varepsilon$ 확률로 현재 Q값 중 최적 행동을 선택한다. Q-table은 상태-행동 쌍에 대해 기대되는 누적 보상을 저장하는 테이블로, 학습이 진행될수록 각 상태에서의 최적 행동이 정의된다[3].

B. 실험 환경 구성

본 논문에서의 실험 배경에 관해 설명하겠다. 본 논문은 실험을 (5 X 5) 정방형 크기의 격자 공간을 실내 환경으로 가정하여 진행하였다. 각 셀은 에이전트가 한 칸씩 이동할 수 있는 공간이며, 두

개의 고정된 장애물이 있다. 에이전트의 시작 위치는 에피소드마다 무작위로 설정되며, 목표 지점은 항상 (4, 4)로 고정되어 있다. 에이전트는 (상, 하, 좌, 우) 중 하나의 행동을 선택할 수 있으며, 각 행동에 대한 보상은 다음과 같다.

1) 목표 지점 도달: +10
2) 장애물 또는 벽에 충돌: −5
3) 정상적인 이동 1회당: −0.5

C. 실험 설정 및 학습 과정

본 논문의 학습은 Q-learning 알고리즘을 사용하였으며, 학습률 α=0.1, 탐험 비율 ε=0.2로 고정하였다. 감쇠율 $\gamma \in \{0.1, 0.3, 0.4, 0.5, 0.7, 0.8, 0.9\}$로 설정하였으며, 각각에 대해 약 1,000개의 에피소드를 통해 학습을 수행하였습니다. 이후 학습된 Q-table에 기반한 정책을 약 100회 테스트하여 1) 평균 보상 2) 이동 거리 3) 충돌률 4) 경로 다양성 등 지표를 측정하였다[4].

D. 실험 결과 및 분석

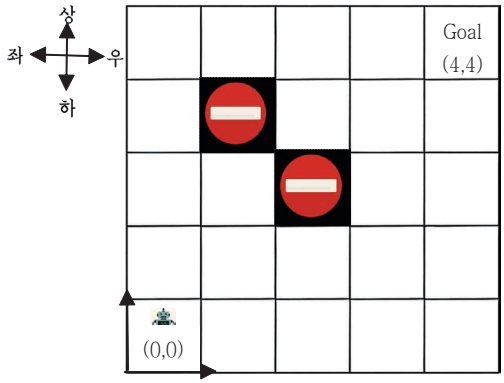


그림 1 실험 배경(5X5)

```
def step(pos, action):
    new = (pos[0] + action[0], pos[1] + action[1])
    if new in obstacles or not (0 <= new[0] < size and 0 <=
        new[1] < size):
        return pos, -5, False
    if new == goal:
        return new, 10, True
    return new, -0.5, False
```

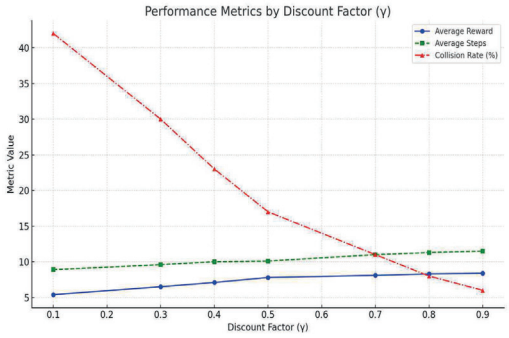
Pseudo-code로, 보상 정의 : 강화학습 환경에서 에이전트가 특정 행동을 했을 때 다음 상태, 보상, 종료 여부(end)를 정의하는 함수이다.

표 1, 본 논문을 통한 실험 결과이다. (경로 다양성은 1~15: 낮음, 16~25: 중간, 26: 높음)으로 측정

감쇠율(γ)	평균 보상	평균 이동 거리	충돌률	경로 수	경로 다양성
0.1	5.4	8.9	42	10	낮음
0.3	6.5	9.6	30	14	낮음
0.4	7.1	10.0	23	18	중간
0.5	7.8	10.1	17	22	중간
0.7	8.1	11.0	11	27	높음
0.8	8.3	11.3	8	31	높음
0.9	8.4	11.5	6	35	높음

본 논문에서의 실험을 바탕으로 한 결과에 관해 설명하겠다.

Table 1에서 보다시피 감쇠율이 낮을수록, 즉 0.1일 때 이동 거리는 짧고 충돌률은 높은 것을 볼 수 있다. 또한, 평균 보상 값이 5.4인 것은 보상받는 때도 있지만, 오히려 실패를 많이 한다는 것을 볼 수 있다. 이러한 관점에서 행동 성향이 단기적 보상에 민감하게 반응하는 것을 볼 수 있다. 반면 감쇠율 값이 클수록, 즉 0.7일 때



Graph 1. Table 1을 시각적으로 보여주는 그래프

이동 거리는 11.0으로 크고 충돌률이 11로 낮은 것을 볼 수 있다. 즉, 에이전트는 장기적인 보상을 고려하여 우회 전략을 취하는 경향을 볼 수 있다. 또한 정상적인 이동의 보상에 대해 에이전트는 짧은 경로를 선호하게 할 수 있게 되었습니다. 그 결과 충돌률이 감소하고 평균 보상이 높고 경로의 다양성 또한 높은 것으로 판단되었다. 경로 이동성은 경로에 있어 실제로 선택하는 유연성으로 정책은 장애물을 통해 충돌이 예상될 때 경로를 유연하게 피할 수 있다. 또한 이동성에 따라 장애물이 있는 환경에서 다른 경로를 선택하여 목표 도달에 쉽다. 이는 정책의 일반화 능력을 높이고 실패율을 줄여 보상을 높일 수 있다. 따라서, 안정적이고 일반화된 정책이 형성되었음을 나타내고 있다[1][3].

Ⅲ. 결론

본 연구는 감쇠율의 설정이 강화학습 기반 경로 계획 정책에 미치는 영향을 정량적으로 분석하였다. 실험 결과, 감쇠율은 단순한 하이퍼파라미터 이상의 의미가 있으며, 경로 계획 정책의 다양성과 행동 전략의 위험 성향을 설계하는 핵심 변수임을 확인할 수 있었다[1][3]. 낮은 감쇠율은 단기 보상 중심의 빠른 경로를 선호하지만, 불안정성이 높았고, 높은 감쇠율은 더 안정적이고 전략적인 정책으로 수렴하였다. 이는 실제 로봇 내비게이션, 사용자 동선 예측, 스마트 공간 제어 등 다양한 응용 분야에서 감쇠율 설정이 실용적인 정책 설계에 미치는 영향을 시사한다[1]. 향후 연구에서는 환경 복잡도를 증가시키거나, 감쇠율의 동적 조절 기법을 적용해 더욱 적응적인 정책을 유도할 수 있는 방향으로 확장할 수 있다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2025년도 SW중심대학사업의 결과로 수행되었음" (2021-0-01389)

참 고 문 헌

[1] ROBOT LEARNING, edited by Jonathan H. Connell and Sridhar Mahadevan, Kluwer, Boston, 1993/1997, pp. 129-132.

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

[3] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

[4] OpenAI. (n.d.). Gym: A toolkit for developing and comparing reinforcement learning algorithms.

[5] Silver, D. (2015). Reinforcement Learning Course (UCL).