밀집된 통신기기의 무선접속 알고리즘에 적용할 수 있는 강화학습 알고리즘에 대한 비교 연구

이선우, 김도원, 윤희준, 이지환, 황건하, 신경섭
상명대학교

fhfgdvd96@naver.com, wlsqur479@gmail.com, jungle126@gmail.com, ji9620@gmail.com, cromany@naver.com, ksshin@smu.ac.kr

A Study on the Reinforcement Algorithm based Medium Access Algorithm in a Densely Deployed Wireless Networks

Lee Seon Woo, Kim Do Won, Yoon Hee Jun, Lee Ji Hwan,
Hwang Geon Ha, and Shin Kyung Seop
Sangmyung University

요 약

본 논문은 기존 강화학습을 활용한 통신 프로토콜 연구에서 주로 다루어진 Q-learning 기반 알고리즘보다 더 기본적인 강화학습 알고리즘인 Temporal-Difference(TD)와 Monte-Carlo(MC)를 통신 알고리즘에 적용하여 강화학습 방식에 따른 네트워크 접속성능의 비교분석을 진행하였다. 실험을 위해 알고리즘은 대용량의 데이터를 작은 단위로 나누어 전송하는 다중 사용자 환경을 가정하였다. 본 논문에서는 TD 와 MC 알고리즘에 동일한 reward 를 주지만 ε-greedy 방식을 사용하는 경우에 ε 값을 다르게 설정하여 성능을 비교하는 실험과 ε 값을 고정하고 reward 를 다르게 하는 실험을 수행하였다. 실험 결과 TD 와 MC 를 활용한 노드 간 통신에 ε값이 작을수록 모두 성능이 향상되는 경향성을 확인하고, 최종 reward 의 차이에 따라 TD 의 성능이 변화가 더 크다는 것을 확인하였다.

Ⅰ. 서론

네트워크 환경이 모바일 기기와 IoT 기기의 증가로 통신량이 급증함에 따라 효율적인 데이터 전송 프로토콜의 필요성이 커져가고 있다. 이에 부응하여 추가적으로 통신시설을 증축하지 않고도 네트워크의 효율성을 높일 수 있는 강화학습을 활용한 통신 프로토콜 알고리즘이 많이 제안되어왔다. 기존 연구에서는 라우팅 프로토콜에 강화학습의 일종인 Q-learning 을 활용하여 노드 경로 선택을 최적화하는 방식으로 네트워크의 효율성을 높이는 연구가 주를 이뤘다. [1][2]

본 논문에서는 통신 알고리즘에 강화학습을 적용하여 서로 다른 노드가 동시에 데이터를 전송하여 충돌하는 상황을 방지하고자 한다. 기존에 연구가 많이 이루어진 노드는 서로의 상황을 모르고 일방적으로 데이터를 전송하므로 통신을 위한 알고리즘으로 model-free prediction 이 적합하다. model-free 알고리즘은 환경에 대한 모델을 세우기 어려운 경우에 적용할 수 있고, 여러 도메인에 동일한 방법을 적용할 수 있다는 장점이 있다. prediction 은 선택한 policy 를 따랐을 때의 value function 을 찾는 문제이며 최적의 해를 찾기보다 value 로의 수렴 여부가 중요하다. 본 논문에서는 기존 연구에서 주로 다루어진 Q-learning 기반 알고리즘 보다 더 기초적인 강화학습 알고리즘 중 Monte-Carlo(MC)와 Temporal-Difference(TD) 두 가지 방식을 통신노드의 미디엄 액세스 프로토콜에 적용하여 성능을 비교함으로써 네트워크의 효율성을 더 높일 수 있는 방안을 비교분석하였다.

Ⅱ. 본론

MC 는 episode 를 진행하며 return 을 저장해두고 episode 가 종료된 이후에 그 결과로 학습한다. return 은 discounted reward 이며 다음과 같다.

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T \quad (1)$$

위의 return 에 따라 policy π 에 대한 value function 은 다음과 같다.

$$v_\pi(s) = E_\pi[G_t \mid S_t = s] \quad (2)$$

Episode 가 진행됨에 따라 다음과 같이 value function 이 업데이트된다.

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t)) \quad (3)$$

MC 는 실행한 후에 통계를 내는 단순한 원리로 episode 가 끝나야만 적용이 가능한데, TD 는 MC 와 다르게 완전하지 않은 episode 에서도 학습이 가능하다. 이 경우 value function 은 다음과 같이 업데이트할 수 있다.

$$V(S_t) \leftarrow V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)) \quad (4)$$

MC 에서의 value function 에서 G_t는 discounted reward 의 합이었지만, TD 에서는 episode 가 완료되어야 하는 G_t 대신에 다음 step 의 예측으로 현재 예측을 업데이트하는 $R_{t+1} + \gamma V(S_{t+1})$ 를 사용한 것을 볼 수 있다.

본 논문에서 진행된 실험은 다음과 같은 환경을 가정하여 설계되었다. 10Mbyte 의 데이터를 1Mbyte 씩 나누어 전송하여 10 개의 slot 으로 이루어진 stage 에 데이터를 다 전송할 경우 한 episode 가 끝난 것으로 가정하고 1Mbyte 를 전송하는 것을 한 step 으로 가정하였다. 크게 MC 알고리즘은 한 episode 가 끝났을 때마다, TD 알고리즘은 한 step 을 수행할 때마다 policy 를 업데이트한다는 차이가 있다.

실험은 총 2 가지 방식으로 진행되었으며 이는 다음과 같다. 실험 1 은 epsilon (ε) -greedy 알고리즘을 활용하여 ε 값이 작을수록 예측 value 값이 가장 큰 행동을 선택할 가능성이 크고 ε 값이 클수록 무작위적으로 행동을 선택한다는 점을 고려하여 ε 값에 따른 randomness 가 각 알고리즘의 성능에 미치는 영향을 분석하기 위한 실험이다. 이를 확인하고자 TD/MC 알고리즘에 0.5, 0.05 의 ε 값을 주어 얼마나 성공적으로 데이터를 전송하였는지를 비교하였다. Fig1 그래프는 실험 1 의 결과를 보여준다.

실험 2 는 TD/MC 알고리즘에 reward 가 미치는 영향을 비교 분석하기 위해 설계되었다. reward 를 R1(한 step 을 수행하고 받는 reward), R2(모든 step 이 끝난 episode 종료 후 받는 reward)로 구분하여 R1=R2 인 경우, R1<R2 인 경우로 reward 값을 조정하여 실험하였다. Fig2 그래프는 실험 2 의 결과를 보여준다.

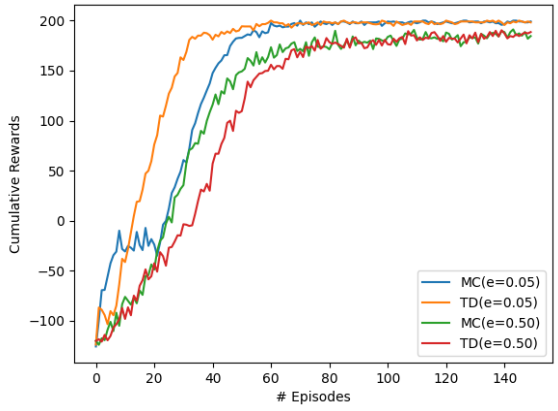

Fig 1. ε 값에 따른 알고리즘의 성능변화

위의 그래프는 MC 와 TD 에 각각 고정된 reward 값에 서로 다른 ε값을 주어 나온 결과이다. ε의 값이 클수록 다음 step 에 random 한 선택을 할 확률이 높아진다. 위의 그래프에서 볼 수 있듯, ε의 값이 작을수록 TD 가 MC 에 비해 더 빨리 수렴하게 되어 유리한 결과를 보여주지만, ε의 값이 커질 경우 그 반대의 결과가 나오게 된다.

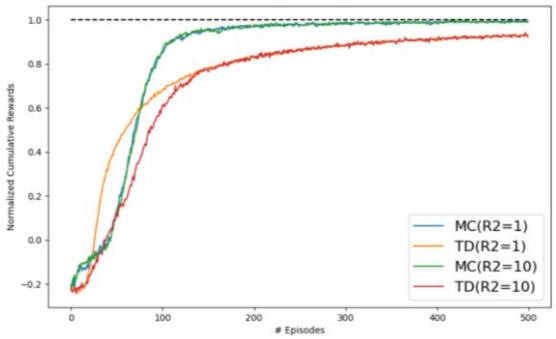

Fig 2. Reward 에 따른 알고리즘의 성능변화

위의 그래프는 reward 값에 따른 MC 와 TD 의 결과이다. 기본적으로 R1 을 1 로 부여하고, R1=R2 일 때와 R1<R2 일 때를 비교하였다. 그래프는 각 unit 이 episode 당 받은 전체 reward 를 서로 다른 random seed 에 대한 평균값으로 표준화하여 나타낸 결과이다. 500 개의 random seed 를 적용한 환경으로, ε-greedy 방식으로 action 을 선택했다. TD 는 step 마다 random 한 확률을 가져 exploration 적인 선택을 하게 했고, MC 는 첫 선택을 해당 step 에 고정적으로 가져가게 하여 TD 에 비해 더 greedy 한 action 선택을 하였다. 위의 그래프에서 볼 수 있듯이, MC 의 경우 reward 의 차이에 큰 영향을 받지 않아 비슷한 그래프를 그리는 경향을 보였고, TD 는 R2 가 커짐에 따라 수렴속도와 변동폭의 안정성이 다소 떨어지는 경향을 보였다.

Ⅲ. 결론

본 논문에서는 MC 와 TD 를 사용한 노드 간 통신에 ε값이 작을수록 빠르게 목표치에 수렴하는 경향성을 확인하였다. ε값이 작을 경우, MC 와 TD 모두 임의적인 선택이 아닌 greedy 한 선택을 더 많이 하게 되므로 ε 값이 큰 경우보다 더 빨리 수렴하게 된다. 반면, reward 의 크기 변화가 MC 에 미치는 영향은 제한적이며 TD 의 경우 Bias 가 생겨 성능에 변동성이 커지는 부작용이 나타날 수 있다. 결과적으로 두 알고리즘 모두에 결정적인 영향을 미치는 요소는 ε값으로, TD 가 더 greedy 한 선택을 할수록 MC 보다 유리해지는 경향을 보인다는 것을 확인하였다.

ACKNOWLEDGMENT

본 연구는 2021 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2021R1F1A1064059).

참 고 문 헌

[1]김태정, 정진우, "주기적 Q-table 업데이트를 활용한 무선 라우팅 프로토콜." 한국통신학회논문지 45.12 (2020): pp. 2099-2105.

[2]김희원, 조호신, "수중 센서 네트워크에서 강화학습 기반의 라우팅 프로토콜." 한국통신학회논문지 45.10 (2020): pp. 1716-1719.