Softmax를 이용한 Q-learning 기반의 패킷 스케줄링

김동현[O], 이태호[*], 이병준[*], 김경태[**], 윤희용[*]
[O*]성균관대학교 정보통신대학 전자전기컴퓨터공학과
[**]성균관대학교 소프트웨어대학 소프트웨어학과
e-mail: {kdh7263, leetaeho, byungjun, youn7147}@skku.edu[O*], kyungtaekim76@gamil.com[**]

Q-learning based packet scheduling using Softmax

Dong-Hyun Kim[O], Tae-Ho Lee[*], Byung-Jun Lee[*], Kyung-Tae Kim[**], Hee-Yong Youn[*]
[O*]Dept. of Electrical and Computer Engineering, Sungkyunkwan University
[**]Dept. of Software, Sungkyunkwan University

● 요 약 ●

본 논문에서는 자원제한적인 IoT 환경에서 스케줄링 정확도 향상을 위해 Softmax를 이용한 Q-learning 기반의 패킷 스케줄링 기법을 제안한다. 기존 Q-learning의 Exploitation과 Exploration의 균형을 유지하기 위해 *e*-greedy 기법이 자주 사용되지만, *e*-greedy는 Exploration 과정에서 최악의 행동이 선택될 수도 있는 문제가 발생한다. 이러한 문제점을 해결하기 위해 본 연구에서는 Softmax를 기반으로 다중 센서 노드 환경에서 데이터 패킷에 대한 Quality of Service (QoS) requirement 정확도를 높이기 위한 연구를 진행한다. 이 때 Temperature 매개변수를 사용하는데, 이는 새로운 정책을 Explore 하기 위한 매개변수이다. 본 논문에서는 시뮬레이션을 통하여 제안된 Softmax를 이용한 Q-learning 기반의 패킷 스케줄링 기법이 기존의 *e*-greedy를 이용한 Q-learning 기법에 비해 스케줄링 정확도 측면에서 우수함을 보인다.

키워드: e-greedy, softmax, 패킷 스케줄링(packet scheduling), 큐러닝(Q-learning)

I. Introduction

최근 기계학습 분야 중 하나로 강화학습에 대한 연구가 활발히 진행되고 있다. 강화학습은 행동심리학을 기반으로 주어진 환경에서 에이전트가 현재의 상태를 파악하고, 실행가능한 행동들 중 보상을 극대화 할 수 있는 행동을 선택하는 방법으로 [1], 2~3년 전 엄청난 이슈가 됐었던 '알파고(AlphaGo)'가 이에 해당한다. 이러한 강화학습과 기존 기계학습 기반 모델과의 가장 큰 차이점은 기존 기계학습 기반 모델은 일일이 사람이 모델링하고 구현해야 했었지만, 강화학습은 스스로 현재의 환경을 파악하여 행동할 수 있다는 점이다. 이를 기반으로 본 연구에서는 강화학습 모델을 IoT 환경에서의 패킷 스케줄링 환경에 적용하여 스케줄링 정확도를 높이는 방법에 대해 연구하였다.

II. Preliminaries

1. Related works

1.1 탐구 전략(Exploration strategy)

e-greedy는 간단하면서도 높은 성능결과를 보이기 때문에 Q-learning의 탐구에 자주 사용된다. 이 때 e는 $0 \leq e \leq 1$의 값으로, 1-*e*의 확률로 기존 정책을 유지(Exploitation)하고 *e*의 확률로 새로운 정책을 탐구(Exploration)한다 [2]. 하지만, 탐구하는 과정에서 무작위로 행동을 선택하기 때문에 Q-learning의 최악 행동(Worst action)을 선택하는 문제가 발생하게 된다.

III. The Proposed Scheme

위와 같은 문제점을 해결하고자 본 논문에서는 Softmax를 이용하여 행동을 선택하는 Q-learning 기법에 대해 연구하였다. Softmax를 기반으로 한 행동 선택은, 상대적인 값을 기반으로 행동이 가중되어 선택된다. Softmax의 수식은 다음과 같다.

$$p(s,a) = \frac{\exp(Q(s,a)/\tau}{\sum_{j=1}^{k}\exp(Q(s,a^j)/\tau}$$

위 수식에서 p(s,a)는 상태 s에서 행동 a를 선택할 확률이며, τ는 탐구의 정도를 조절하는 Temperature 매개변수이다. 본 연구의 정확

성을 검증하기 위해 본 논문에서는 3개의 큐와 1개의 게이트웨이로 구성된 네트워크 환경에서, 제안된 Softmax를 이용한 패킷 스케줄링 기법과 기존 *e*-greedy 기법을 이용한 패킷 스케줄링 기법의 정확도를 비교하였다.

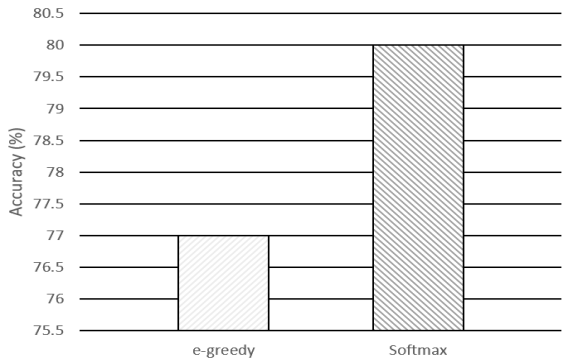

Fig. 1. Simulation results

시뮬레이션 결과 제안된 Softmax를 이용한 패킷 스케줄링 기법의 정확도가 기존 *e*-greedy 기법에 비해 향상된 결과를 보였다.

IV. Conclusions

본 연구에서는 다양한 기계학습 분야 중 하나인 강화학습을 이용하여, 패킷 스케줄링 정확도를 향상시키기 위해 Softmax를 이용한 Q-learning 기법에 대한 연구를 진행하였다. Softmax는 Q-learning의 학습에 사용된 Q-values에 대한 상대적으로 가중된 확률을 기반으로 행동을 선택하는 방법으로, 시뮬레이션 결과 기존 *e*-greedy를 이용한 패킷 스케줄링 기법에 비해 향상된 정확도를 보였다. 향후 연구계획으로 보다 정확도를 향상시키기 위한 연구를 진행할 예정이다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 정보통신·방송연구 개발 사업(No. 2016 -0-00133, 초연결 IoT 노드의 군집 지능화를 통한 Edge Computing 핵심 기술 연구), SW중심대학지원사업(2015-0-00914), 한국연구재단 기초연구사업 (No.2016R1A6A3A11931385, 실시간 공공안전 서비스를 위한 소프트웨어 정의 무선 센서 네트워크 핵심기술 연구, 2017R1A2B2009095, 실시간 스트림 데이터 처리 및 Multi-connectivity를 지원하는 SDN 기반 WSN 핵심 기술 연구), BK21PLUS 사업의 일환으로 수행되었음.

REFERENCES

[1] https://ko.wikipedia.org/wiki/강화_학습.

[2] H. Ferra, et al, “Applying Reinforcement Learning to Packet Scheduling in Routers,” In Proceeding of the Fifteenth Innovative Applications of Artificial Intelligence Conference, pp.79-84, 2003.