UCB를 이용한 강화학습 패킷 스케줄링

김동현[O], 김민우[*], 이병준[*], 김경태[**], 윤희용[*]
[O*]성균관대학교 정보통신대학 전자전기컴퓨터공학과
[**]성균관대학교 소프트웨어대학 소프트웨어학과
e-mail: {kdh7263, kimmw95, byungjun, youn7147}@skku.edu[O*], kyungtaekim76@gamil.com[**]

Reinforcement learning packet scheduling using UCB

Dong-Hyun Kim[*], Min-Woo Kim[*], 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 Upper Confidence Bound (UCB)를 이용한 효율적인 패킷 스케줄링 기법을 제안한다. 기존 *e*-greedy 등 강화학습의 보상을 극대화 할 수 있는 행동을 선택하는 것과 다르게, 제안된 UCB를 이용한 강화학습 패킷 스케줄링 기법은 각 상태에서 행동을 선택한 횟수를 추가적으로 고려한다. 이는 보다 효율적인 강화학습의 탐구(Exploration)를 가능케 한다. 본 논문에서는 컴퓨터 시뮬레이션을 통하여 제안하는 UCB를 이용한 강화학습 패킷 스케줄링 기법이 기존의 *e*-greedy 및 softmax를 기반으로 한 패킷 스케줄링 기법에 비해 정확도 측면에서 향상된 정확도를 보인다.

키워드: softmax, Upper confidence bound, 탐구(exploration), 큐러닝(Q-learning), 패킷 스케줄링(packet scheduling)

I. Introduction

최근 하드웨어 장치 및 네트워크 기술의 발달로, 다수의 센서 노드로 구성된 IoT 환경에 대한 연구가 활발히 진행되고 있다. 이러한 이기종 IoT 환경은 스마트 홈, 환경 모니터링 등 다양한 애플리케이션을 포함하고 있으며 방대하고 다양한 종류의 데이터를 생성한다. 이중에서도 특히 실시간 데이터가 통신 데이터의 대부분을 차지하기 때문에 사용자에게 품질있는 서비스를 제공하기 위해서는 Quality of Service (QoS)를 고려해야 한다. 여기서 QoS 요구조건이란 응용, 사용자, 그리고 데이터의 흐름 등에 우선 순위를 부여하여 일정 수준 이상의 성능을 보장하기 위해 필요한 요소이다 [1]. 본 논문에서는 다수의 센서노드 환경에서 수집되는 데이터 패킷의 QoS 요구조건을 고려하기 위해 강화학습을 이용하는데, 다양한 강화학습 방법 중 Q-learning을 이용한 UCB 행동 선택 기법을 사용한다.

II. Preliminaries

1. Related works

1.1 탐구 전략(Exploration strategy)

Softmax 기법은 Q-learning의 학습에 사용되는 Q-values에 대해 0과 1 사이의 상대적인 확률을 부여하고, 할당된 확률을 기반으로 행동을 선택하는 방법이다. Softmax의 결과 값의 합은 항상 1이다. 예를 들어, Q-values에 대한 상대적인 확률이 0.5, 0.3, 0.2 로 계산되었다면, 각 확률로 해당하는 행동을 선택한다. 하지만 Softmax 기법 또한 항상 최적의 행동을 선택하지 못한다.

III. The Proposed Scheme

본 논문에서는 Q-learning을 이용한 패킷 스케줄링 기법의 QoS 요구조건 정확도를 높이기 위해 UCB를 사용한다. UCB는 상태 *s*에 따른 행동 *a*가 선택된 횟수를 고려하는 방법이다.

$$a_t = \arg max\left(Q(s,a) + \sqrt{\frac{2\log(t)}{N(s,a)}}\right)$$

위 수식에서 N(s,a)는 상태 *s*에서 행동 *a*가 선택된 횟수이며, *t*는 시간이다. 즉 UCB를 이용한 Q-learning 패킷 스케줄링 기법은 한 번도 선택되지 않은 행동에 대해 높은 가중치를 부여하여 선택하고, 자주 선택된 행동에 낮은 가중치를 부여하는 방법이다. 위 연구 내용을 기반으로 기존 softmax 기법과 정확도 측면에서의 성능을 비교실험 하였다.

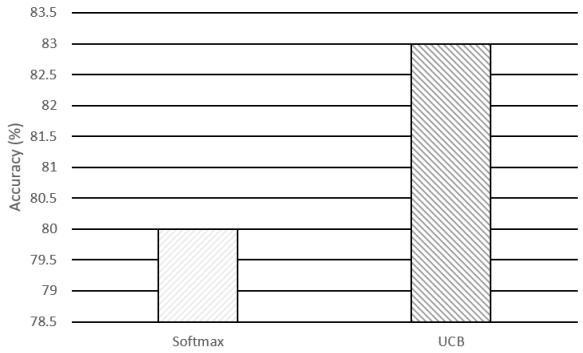

Fig. 1. Simulation results

위 그림은 제안된 UCB와 Softmax 기법을 이용한 행동 선택의 정확성 결과로 제안된 UCB 기법의 정확도가 더 높은 결과를 보였다.

IV. Conclusions

본 논문에서는 강화학습에서의 다양한 행동 선택 방법 중 UCB 기법을 이용한 패킷 스케줄링에 대한 연구를 진행하였다. 강화학습에서의 가장 큰 요점은 Exploitation과 Exploration의 균형을 맞추는 것인데, UCB는 선택된 횟수가 적은 행동을 선택함으로써 Exploration한다. 시뮬레이션 결과 제안된 UCB를 이용한 Q-learning의 패킷 스케줄링 기법이 기존 Softmax 및 *e*-greedy 기법에 비해 정확도 측면에서 성능이 우수함을 보였다.

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/QoS

[2] Kdhong, "An Efficient Dynamic Workload Balancing