감독 지식을 융합하는 강화 학습 기법을 사용하는 셀룰러 네트워크에서의 동적 채널 할당 기법

김성완° 장형수
서강대학교 컴퓨터공학과
inaina21@sogang.ac.kr, hschang@sogang.ac.kr

A Dynamic Channel Assignment Method in Cellular Networks Using Reinforcement Learning Method that Combines Supervised Knowledge

Sungwan Kim° Hyeongsoo Chang
Department of Compute Science and Engineering, Sogang University

1. 서 론

강화 학습(reinforcement learning, 이하 RL)[1] 알고리즘은 마르코프 의사결정 과정(Markov Decision Process, 이하 MDP)[1] 모델의 상태전이 함수와 보상 함수를 알지 못해도 최적 정책을 학습할 수 있다는 장점이 있는 반면, 그 최적 정책에 느리게 수렴한다는 단점이 있다. 강화 학습의 수렴 속도를 향상시키려는 연구는 현재까지 활발하게 이루어지고 있다. 최근에는 "감독" 지식(supervised knowledge)을 강화 학습의 과정에 융합하여 수렴 속도를 향상시키려는 연구가 진행되어 왔다. 특히 Chang [2] 은 "potential-based reinforcement function"을 이용하여 감독 지식을 융합한 학습 기법(이하 potential-based RL 기법)을 제시하였다. 이는 다수 학습(multiple learning)들과 expert advice들을 감독 지식으로 강화 학습에 융합하는 것을 가능하게 했고, 이전의 연구들과 다르게 그 이론적인 최적 정책으로의 수렴성을 확립하였다.

본 논문에서는 potential-based RL 기법을 셀룰러 네트워크에서의 채널 할당 문제[4]에 적용한다. 셀룰러 네트워크상에서 사용할 수 있는 채널의 수는 제한되어 있기 때문에, 각 셀에 대한 효율적인 채널 할당 문제는 셀룰러 네트워크 디자인의 중요한 이슈로 여겨져 왔다. 잘 알려진 채널 할당 기법인 MAXAVAIL[3]을 expert로 사용하는 potential-based RL 기반의 동적 채널 할당 기법은 FCA, Q-Learning based dynamic channel assignment[4], MAXAVAIL 채널 할당 기법에 비해 보다 효율적으로 채널을 할당한다. 또한, 강화 학습 기법간의 실험적인 성능 비교를 통하여 potential-based RL 기법이 Q-learning에 비해 최적 정책에 더 빠르게 수렴함을 보인다.

2. 본 론

2.1 Potential-based RL 기법

Potential-based RL 기법은 SARSA(0) 학습 기법[1]을 사용하는 기본 에이전트(base-agent)와 Q-learning, SARSA(λ) 등의 학습 기법을 사용하는 하나 이상의 서브 에이전트(subagent)로 구성되며, 여기에 감독 지식을 제공하는 다수의 expert가 추가될 수 있다. 서브에이전트들의 다수 학습(multiple learning)들에 의한 Q 값[1]의 추정치와, expert들에 의한 감독 지식은 potential-based reinforcement function[6]에 의해서 기본 에이전트의 학습에 영향을 미친다. MDP $M = (X,A,P,R)$이 주어졌다고 할 때 Potenrial-based RL 기법에서 Q 값의 업데이트 공식은 다음과 같다.

$$Q_{t+1}(x_t, a_t) \leftarrow Q_t(x_t, a_t) + \alpha_t(x_t, a_t)[R(x_t,a_t,x_{t+1}) + \gamma\Phi(x_{t+1}) - \Phi(x_t) + \gamma Q_t(x_{t+1}, a_{t+1}) - Q_t(x_t, a_t)]$$

이는 SARSA(0)의 업데이트 공식[1]에 potential-based reinforcement function Φ를 추가한 것이며, m개의 서브에이전트와 l개의 expert들이 있다고 할 때 Φ는 다음과 같이 정의된다[2].

$$\Phi(x_t;t_1, \ldots, t_m,s_1,\ldots,s_l) = \sum_{a\in A}\left(\frac{1}{m}\sum_{i=1}^{m} Q_{t_i}^{i}(x_t, a_t) \times \theta(x_t, a;t_1, \ldots, t_m,s_1,\ldots,s_l)\right)$$

$Q_{t_i}^{i}$-함수는 서브에이전트 i가 자신의 학습 기법을 사용하여 학습하는 Q 함수의 t_i에서의 추정 값을 말한다. 각각의 expert들은 행동들의 집합 A에 대한 확률 분포의 형태로 기본 에이전트에 expert advice들을 제시한다고 하자. 여기서 $\theta(x_t, a;t_1, \ldots, t_m,s_1,\ldots,s_l)$는 다음과 같이 주어진다.

※ 이 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2007-000-10511-0).

$$\theta(x_t, a; t_1, \ldots, t_m, s_1, \ldots, s_l) = \frac{\sum_{i=1}^{m} I\left(a \in \arg\max_{b \in A} Q_{t_i}^i(x_t, b)\right)}{\sum_{a' \in A} \sum_{i=1}^{m} I\left(a' \in \arg\max_{b \in A} Q_{t_i}^i(x_t, b)\right)} \times \frac{\sum_{i=1}^{l} \rho_{s_k}^k(x_t, a)}{\sum_{a' \in A} \sum_{i=1}^{l} \rho_{s_k}^k(x_t, a')}$$

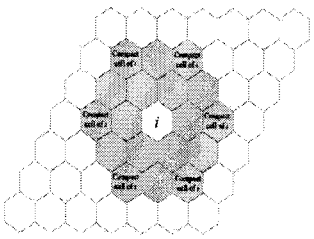
$\rho_{s_k}^k(x_t, a)$는 전문가 s_k가 상태 x_t에서 행동 a를 택할 확률이다. potential function Φ 를 통하여 m개의 서브에이전트들 각각의 Q 함수의 추정 값들과 expert advice들을 기본 에이전트의 업데이트 공식에 반영할 수 있으며, $t \to \infty$에 따라 기본 에이전트 SARSA(0)에 의하여 학습된 policy는 최적 정책(optimal policy)에 수렴하게 된다[2].

2.2 Channel Allocation in Cellular Network

Cellular network의 채널 할당 문제에 위의 Potential-based RL 기법을 적용하였다. 이는 Q-learning based Dynamic Channel Allocation(이하 DCA)[4]를 potential-based RL 기반 DCA로 확장시킨 것으로서, Q-learning을 사용하는 서브에이전트와 MAXAVAIL[3] 을 사용하는 expert로 구성되어 있다. MAXAVAIL은 Cellular network의 채널 할당 문제를 해결하는 잘 알려진 알고리즘이다.

3. 결 론

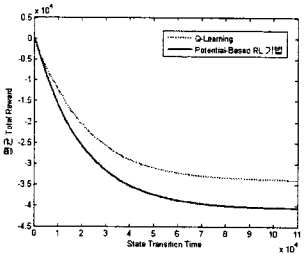
실험은 [그림1]과 같이 7×7 개의 육각형의 cell들로 구성된, 70개의 채널을 가진 셀룰러 네트워크를 가정하였으며, 기타 환경 설정은 [4]와 동일하게 진행하였다.



[그림 1] Cellular Network

Traffic Load	Q-learning	FCA	Maxavail	Potential-Based RL
5	0.139%	1.750%	0.086%	0.102%
6	1.766%	4.101%	1.344%	1.090%
7	5.708%	7.570%	4.673%	4.590%
8	10.235%	11.724%	10.004%	9.433%
9	15.076%	16.409%	15.014%	14.303%
10	20.216%	21.196%	19.824%	18.983%

[표 1] Blocking Probability 비교



[그림 3] 수렴 과정 비교

Cellular network에서의 채널 할당 알고리즘의 성능 척도인 blocking probability에 대한 실험을 위해, FCA와 Maxavail은 5시간동안 각각의 채널 할당 기법을 수행한 뒤 blocking probability를 측정하였고, Q-learning-based DCA와 potential-based RL을 적용한 DCA는 15시간동안 각각의 정책을 학습한 뒤, 해당 정책을 사용하여 5시간동안 채널 할당을 수행한 뒤 blocking probability를 측정하였다. [표 1]에서 볼 수 있듯이, potential-based RL을 사용한 DCA의 경우 나머지 세 채널 할당 기법보다 낮은 blocking probability를 보임을 알 수 있다. 이는 네 개의 채널 할당 기법들 중 potential-based RL 기법이 가장 효율적으로 채널을 할당한다는 것을 뜻하며, Maxavail을 expert로 사용한 potential-based RL이 기존의 강화 학습의 성능을 향상시켰음을 실험적으로 증명한 것이다. [그림 3]은 Potential-based RL 기법을 사용한 DCA와 Q-learning-based DCA의 수렴 과정을 비교한 그래프이다. 11×10^4 번의 상태 전이를 거친 결과 potential-based RL 기법이 Q-learning에 비해서 낮은 평균 total reward값을 보였다. 무한한 시간동안 학습할 경우 Q-learning과 potential-based RL 기법은 모두 최적 정책에 수렴하며 같은 평균 total reward를 갖게 된다[2]. 따라서 같은 시간동안 동일한 조건에서 학습하였을 때 potential-based RL 기법이 Q-learning에 비하여 낮은 평균 total reward값을 보인다는 것은 potential-based RL 기법이 최적 정책에 Q-learning보다 더 빠르게 수렴한다는 것을 뜻한다. 즉 이 실험을 통하여 potential-based RL 기법이 기존의 강화 학습의 성능을 향상시켰음을 확인할 수 있다.

참고문헌

[1] R. Sutton and A. Barto, *Reinforcement Learning*. MIT Press, 2000.

[2] H. S. Chang, "Reinforcement Learning with Supervision by Combining Multiple Learnings and Expert Advices", in *Proc. of the 2006 American Control Conference*, June, 2006, pp. 4159-4164.

[3] Sivarajan, K.N.; McEliece, R.J.; Ketchum, J.W., "Dynamic channel assignment in cellular radio," Vehicular Technology Conference, 1990 IEEE 40th , vol., no., pp.631-637, 6-9 May 1990

[4] Junhong Nie; Haykin, S., "A dynamic channel assignment policy through Q-learning," Neural Networks, IEEE Transactions on , vol.10, no.6, pp.1443-1455, Nov 1999