자율 이동 로봇의 주행을 위한 영역 기반 Q-learning

Region-based Q-learning For Autonomous Mobile Robot Navigation

○ 차 종 환, 공 성 학, 서 일 홍

한양대학교 전자공학과 (Tel : +82-31-408-5802; Fax : +82-31-408-5803;
E-mail: ihsuh@email.hanyang.ac.kr)

Abstract : Q-learning, based on discrete state and action space, is a most widely used reinforcement learning. However, this requires a lot of memory and much time for learning all actions of each state when it is applied to a real mobile robot navigation using continuous state and action space. Region-based Q-learning is a reinforcement learning method that estimates action values of real state by using triangular-type action distribution model and relationship with its neighboring state which was defined and learned before. This paper proposes a new Region-based Q-learning which uses a reward assigned only when the agent reached the target, and get out of the local optimal path with adjustment of random action rate. If this is applied to mobile robot navigation, less memory can be used and robot can move smoothly, and optimal solution can be learned fast. To show the validity of our method, computer simulations are illustrated.

Keywords : RQ-learning, neighboring state, action distribution model, mobile robot, navigation

1. 서론

자율 이동 로봇을 제어하는 방법으로 강화 학습이 많이 사용되는데 이것은 알려지지 환경에서 행동과 보답을 주고 받으며 임의의 상태에서 가장 적합한 행위를 학습하는 방법으로 trial-and-error 책략에만 의존한다.[1]

이러한 방법들 중 Q-learning은 가장 널리 사용되는 방법들 중 하나로 이 학습법은 현재 상태에서의 행위를 미래 행위들로부터 얻게 되는 총 보답을 예측하는 행위값에 대응시키도록 하는 행위 함수를 학습하는 방법이다[3]. 그런데, Q-learning은 오차를 줄이기 위해서 많은 기억 공간과 매우 긴 학습 시간이 필요하다는 점 때문에 실제 환경에 적용하기는 쉽지 않다. 이러한 단점을 보완하기 위해 여러 가지 연구가 진행되어 왔다. 예를 들면, Q-table의 크기를 가변하는 방법과 보답을 실수로 얻는 방법, 그리고 행위값의 초기화나 좋은 탐색 책략을 미리 알려주는 방법이다[4,5,6,7].

하지만, 강화 학습의 대부분의 연구들은 불연속 상태공간과 불연속 행위공간을 기반으로 이루어졌기 때문에 실제 자율 이동 로봇의 주행에 적용시키려면 연속 상태 공간의 연속 행위에 대해 학습하여야 하므로 많은 기억공간과 긴 학습 시간이 필요하며 또한 행위 가 불연속이라는 제한을 갖게 되고, 정의된 상태들이 정확한 현재 상태를 나타낼 수 없기 때문에 오차가 존재하게 된다[8].

이러한 제약은 RQ-learning을 사용함으로써 줄일 수 있게 되는데 기존의 RQ-learning은 보답을 얻기 위해서는 모든 상태에서 목표 지점까지와의 거리를 안다는 가정이 필요하였고, 또, 지역적 최적해를 벗어나지 못 할 수도 있다는 단점이 있다[2]. 본 논문에서는 이러한 점을 보완하여 자율 이동 로봇에 적용하여 보고, 컴퓨터 모의 실험을 통하여 Q-learning을 적용하였을 때의 결과와 비교함으로써 제안한 RQ-learning의 효과적인 점을 검증하여 본다.

2. Q-learning과 RQ-learning

2.1 Q-learning

Q-learning은 대표적인 off-policy 강화학습으로 주어진 환경과 상호작용을 하며 최적의 행위 함수를 학습하는 방법이다. 임의의 상태에서 최적의 행위책략(policy) $\pi^*(s_t)$는 행위 함수로부터 결정된다.

$$\pi^*(s_t) = \arg\max_a Q(s_t, a) \tag{1}$$

그리고, 행위 함수는 다음과 같이 갱신된다.

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left(r_{t+1} + \gamma \max_a \{Q(s_{t+1},a)\} - Q(s_t,a_t)\right) \tag{2}$$

Q-learning의 알고리즘을 간략히 소개하면 다음과 같다.

[Q-learning 알고리즘]

1. Q-테이블 및 각 파라미터 (α,γ)들을 알맞게 초기화한다.
2. 다음 내용을 반복한다.
   1) Q-테이블로부터 행위를 결정한다. 이때 임의의 비율로 랜덤 행위를 만들어 주어야 한다.
   2) 행위에 대한 다음 상태와 보답을 얻는다.
   3) 현재 상태의 행위값을 갱신한다.
   4) 행위책략을 갱신한다.

2.2 RQ-learning

기존의 RQ-learning은 모든 상태에서 목표점까지의 거리를 안다는 가정이 필요하고, 지역적 최적해를 벗어나지 못 할 수도 있기 때문에 이를 보완한 RQ-learning을 제안한다.