Proceedings of the 11th
KACC, October 1996

동적 환경에서 강화학습을 이용한 다중이동로봇의 제어

Reinforcement Learning for Multi Mobile Robot Control in The Dynamic Environments

°김도윤, 정명진

한국과학기술원 전기 및 전자공학과
(Tel:042-869-5429; Fax: 042-869-3410; E-mail: nice@donghae.kaist.ac.kr)

Abstracts

Realiazation of autonomous agents that organize their own internal structure in order to behave adequately with respect to their goals and the world is the ultimate goal of AI and Robotics. Reinforcement learning gas recently been receiving increased attention as a method for robot learning with little or no a priori knowledge and higher capability of reactive and adaptive behaviors. In this paper ,we present a method of reinforcement learning by which a multi robots learn to move to goal. The results of computer simulations are given

Keyword Machine learning, Reinforcement learning, Q learning, mobile robots

1. 서론

이동로봇의 항법에 관한 연구는 로봇의 위치를 정확히 알아내는 위치추정에 관한 연구와 목적지를 찾는 경로계획, 그리고 목적지까지 이동하기 위한 경로제어로 나눌 수 있다. 산업 현장에서 실용화되고 있는 대부분의 이동로봇은 바닥에 설치되어 있는 고정궤도를 따라 운행하는 방식으로서 이 방법은 항상 주어진 경로만을 주행하므로 경로계획과 경로제어문제가 간단히 해결된다. 그러나 돌발적인 상황이 발생하여 경로상에 장애물이 나타나거나 경로를 이탈하는 경우에 대한 대처능력이 없어 동적으로 상황이 변하는 공간에는 적용이 불가능하다. 최근에는 산업현장에서 벗어나 일반적인 공공장소에서 동작하는 서비스로봇이나 극한지역에서 불확실한 환경에서 동작하는 로봇에 대한 연구가 활발한데 이러한 로봇은 기존에 알고 있는 환경에 대한 정보가 바뀌더라도 바뀐 환경에 적응할 수 있는 학습능력이 필요하게 된다.

강화학습은 객체가 어떠한 행위를 했을 때 행위에 대한 보답(reward)을 받아서 보상값을 좀 더 높이는 방향으로 계속 행위를 이끌어 가는 방법으로서 선험적인 지식이 필요 없이 학습을 할 수 있고 간단한 알고리즘으로 환경에 적응성이 뛰어난 행위를 생성하여 높은 반응을 보인다. 이러한 강화학습 방법 중에 Q-learning은 주어진 공간에서 취할 수 있는 상태를 불연속적인 상태공간으로 설정하고 각 상태가 취할 수 있는 행위를 설정한 후 현재 상태에서 행동을 취했을 때 기대되는 값을 학습하여 목표상태에 도달하기 위한 최적의 행위공간을 구한다. 특히 Q-learning의 계산량을 줄인 Temporal Difference Q-learning(TDQ)과 prioritized sweeping technique는 최근 집중적으로 연구되고 있다.

본 논문에서는 Q-learning을 이용하여 주어진 환경에서 목표점을 찾고 환경이 학습한 경우와 달리 움직이는 물체가 경로상에 존재하여도 목표점에 도달하는 방법과 여러 대의 로봇에 적용하는 방법에 대해 제안한다. 2장에서는 연구동향을 소개하고 3장에서는 학습방법으로 사용된 Q-learning에 대해 간단히 소개한다. 4장에서는 환경을 학습한 후 그 환경에서 움직이는 물체가 있는 경우의 경로제어, 여러 대의 로봇이 있는 경우의 경로제어방법을 제안하고 5장에서는 모의실험의 결과를 제시하였다.

2. 연구동향

여러 대의 이동로봇이 같은 공간에서 움직이는 연구분야는 로봇과 로봇이 통신을 할 수 있는 경우와 통신을 할 수 없는 경우로 나누어 생각할 수 있다. 이 때 통신이 되지 않는 경우는 algorithm의 중요성이 더욱 강조된다. 이전에 제안되었던 방법에는 움직이는 물체에 우선 순위를 부여하여 경로를 생성하는 방법과 물체의 속도와 궤적을 알아내어 물체의 경로를 생성하여 피해 가는 방법, 가상 임피던스(virtual-impedance)를 적용하여 경로를 생성하는 방법이 있었다. Causse[4]는 병원에서 필요한 도구를 이동하는 로봇들을 제어하는데 있어 중앙에 모든 로봇을 감시하는 모니터링 시스템을 구축하고 로봇의 충돌이 예상되는 지역에 로봇이 진입하는 경우 다른 로봇은 그 지역에 진입을 금지하는 방법을 사용하였고 Tsubouchi[8]은 주어진 공간상에서 움직이는 물체의 속도는 항상 일정하다고 가정하고 물체는 원형으로 가정한 후 2차원 평면(x,y)에 수직으로 시간축을 가정한 3차원 공간(x,y,t)에서 물체의 위치를 예측한다. 이러한 예측을 바탕으로 로봇에게 필요한 경로는 직진과 호로 이루어진 곡선으로 구성된다. 움직이는 물체의 위치를 예측하기 위해 경로탐색은 로봇이 목적지를 향해 움직이는 동안 계속 반복된다. 이러한 방법들은 모두

다른 로봇의 움직임을 알고 있어야 한다는 가정이 있어 일반적인 환경에서 적용이 불가능하다. 최근에는 다른 로봇의 움직임을 알 필요가 없는 알고리즘 개발에 관심을 가지고 있다..

3. 강화학습의 소개

강화학습은 문제는 경우에 따라 일어나는 보상값으로 미지의 환경에 어떻게 적응시킬 것인가로 표현할 수 있다. 로봇은 환경에서 학습해야 할 상태의 종류를 미리 알고 있다고 가정하에 크게 두 가지 방법으로 구현할 수 있다. 첫 번째는 모델기반제작(model-based design)으로 로봇이 움직일 환경을 이미 알고있다는 가정하에 환경의 모델 M과 유용성함수 U(utility function)로 정의된다. 이 경우는 잘 알려진 dynamic programming에 의해 최적의 정책을 결정할 수 있을 것이다. 다른 방법으론 환경을 모르는 상황에서 환경의 모델 M이 필요 없는 방법으로서(model-free approach) 행위값(action-value)을 나타내는 Q 함수를 사용하는 Q-learning이 있다. 본 논문에서는 환경에 대한 정보가 없다는 가정에서 출발하기 때문에 강화학습중에서 Q-learning에 대해 살펴본다.

3.1 Q-learning

로봇은 명확한 환경의 상태집합 S를 구별할 수 있고 그 환경에서 행위집합 A에서 행위를 취할 수 있다고 가정한다. 환경은 현재상태에서 로봇이 선택해야할 행동을 통계적인 방법으로 결정할 수 있는 Markov process로 정의된다. 현재의 상태-행위쌍인 (s,a)에서 다음 상태인 s'으로 이동할 확률을 상태전이확률이라 정의하고 $T(s,a,s')$으로 놓으며 각각의 상태-행위쌍인 (s,a)에서 보답인 $r(s,a)$이 정의된다.

일반적인 강화학습문제는 시간이 지남에 따라 감소하는 보답의 합을 최대화하는 정책(policy)을 찾는 문제로 생각할 수 있다. 정책 f 는 관련 있는 S와 A를 연결하는 기능을 하게 된다. 이러한 보답의 합은 다음과 같이 정의된다.

$$\sum_{n=0}^{\infty}\gamma^{n}r_{t+n} \qquad (1)$$

로봇이 환경의 상태를 감지하여 행동하는 시간을 단계(step)라고 가정하면 r_t는 t단계에서 로봇이 상태 s에서 정책 f를 실행했을 경우 받는 보답으로 정의된다. γ는 시간에 따라 감소하는 감쇠상수이며 보답이 과거의 정책에 얼마만큼의 영향을 미치는가를 나타내며 1보다 작은 값을 취한다. 그러나 로봇은 움직이는 환경에 대한 정보가 없기 때문에 상태전이확률을 알 수 없으며 정책 또한 결정할 수 없기 때문에 Wakin은 Q-learning을 개발하였다. $Q(s,a)$는 어떤 상태 s에서 행위 a를 선택했을 때 최적의 정책을 선택하기 위한 행위값을 나타내며 정의는 다음과 같다.

$$Q(s,a) = r(s,a) + \gamma \sum_{s' \in S} T(s,a,s') \max_{a' \in A} Q(s',a') \qquad (2)$$

여기서 초기에 상태전이확률 T와 보답 r에 대해 알지 못하기 때문에 로봇은 온라인(on-line)으로 Q값을 추정한다. 초기에 $Q(s,a)$는 임의의 값으로 초기화를 하고 (일반적으로 0), 매 단계에서 행위를 선택하여 취한 후 Q값은 다음과 같은 식으로 수정된다.

$$Q(s,a) \Leftarrow (1-\alpha)Q(s,a) + \alpha(r(s,a) + \gamma \max_{a' \in A} Q(s'a')) \qquad (3)$$

위 식에서 r은 상태 s에서 행위 a를 취했을 때 받는 보답이며, s'은 다음상태를 나타내며 α는 학습률을 나타낸다.

4. 동적환경에서 Q-learning

주어진 환경에서 목적지가 어디에 있는지 알고 있다면 목적지와 현재 위치와의 거리의 절대치를 사용하여 Q-learning에서 필요한 보답을 주면 될 것이다. 하지만 목적지가 어디에 위치하는지 모르고 목적지에 도달했을 때만 목적지인지 알 수 있는 상황에서는 이러한 방법을 사용할 수 없다. 따라서 상태 s 에서 행위 a를 취하여 상태가 s'으로 바뀌었을 때 s'가 목적지가 아니라면 정해진 일정한 음수의 값을 보답으로 주었다. 따라서 목적지를 찾기 전까지는 Q값은 모두 음수가 되며, 목표점을 찾은 후에야 목표점 부근에서부터 출발점까지 양수값이 확산된다. 모의실험에 사용한 환경과 가정은 다음과 같다.

상태 : 가로 16, 세로 16 (전체 256의 상태공간)
행위 : 전,우,좌,우와 각각의 대각선 방향, 그리고 정지하는 경우로서 총 9개의 행위
출발점 : (1,1) 목적지 : (16,16)
보답 : 목적지(1), 물체를 감지한 경우(-0.02)
환경 : box-canyon

가정

- 초기에 환경을 인식하는 단계에선 움직이는 물체는 없다.
- 로봇이 목적지를 찾으면 출발점에서 다시 학습을 시작한다.
- 로봇이 움직이는 물체 앞까지 갔을 때만 물체를 인식할 수 있다.
- 움직이는 물체는 로봇이 목표점을 향해 움직이는 경로를 막는 위치로 항상 이동한다.
- 움직이는 물체와 로봇이 동일한 상태공간을 공유하지 않는다.

4.1 미지의 환경에서의 학습과 탐색

로봇이 알지 못하는 환경을 학습할 때 가장 큰 문제는 최적의 경로를 얻기 위해 학습이 더 필요한지 결정하는 문제이다. blind search 방법으로 각 상태에서 취할 수 있는 모든 행위를 취해보기 전에는 완벽한 최적의 경로를 얻는다는 보장은 없다. 그러나 모든 경로를 탐색하는 경우 상태공간이 증가함에 따라 계산량은 폭발적으로 증가하여 효율성이 떨어진다. 학습한 경로 이외의 상태를 탐색하는 확률로 Q값에 기반을 둔 볼츠만 분포를 갖는 확률적인 행위를 사용할 수 있다. 주어진 상태 s에서 우리는 다음과 같은 확률 분포에 따라 행위를 선택하게 된다.

$$\frac{e^{Q(a,s)/T}}{\sum_{a \in A} e^{Q(a,s)/T}} \qquad (4)$$

온도매개변수 T를 이용하여 탐사시간을 제어할 수 있다. 다른 방법으론 이전의 행위값의 분산을 이용한 2차 정보를 생각해 볼 수 있다. 이는 기존의 학습된 행동에 기인한 data들로서 좀 더 효율적이다. 이 때는 각 행위 a_i에 대한 통계값을 다음과 같이 저장한다.

w_i : 성공 횟수

n_i : 시도 횟수

각 상태에서 행위는 100*(1- α)%의 신뢰구간을 가지고 가장 높은 확률을 선택한다. 이 때 α가 작을 수록 많이 탐색하게 된다. 본 논문에서는 이 방법을 사용하였다.

Fig.1은 box canyon 문제에 적용한 모의실험결과이다. 초기에는 골짜기 부분에 빠져 이리저리 탐색을 시도하다가 시도횟수가 늘어남에 따라 계곡을 빠져나와 다른 방향으로 탐색을 하여 목적지를 찾게된다. 0 보다 큰 값으로 표시된 부분은 목적지까지 이동하는데 성공한 상태집합을 나타내고 음수의 값을 갖는 부분은 실패한 상태집합을 나타낸다.

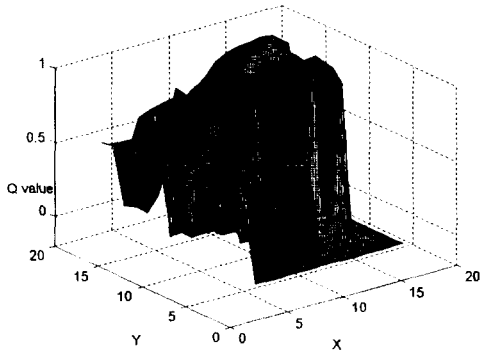

그림 1. 25번 반복한 후 Q값
Fig 1. Q-value (25 iteration)

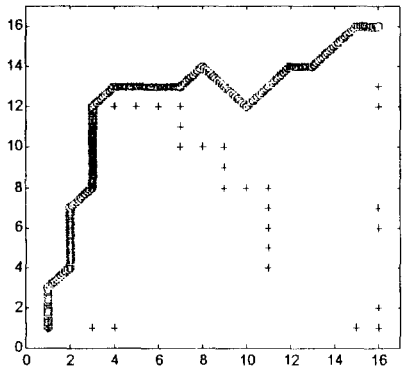

그림 2. 학습한 값으로 주행
Fig. 2 Path planning

4.2 움직이는 물체가 있는 동적 환경에서 학습과 방법

환경에 대해 학습을 한 후 $Q(s,a)$에서 목적지까지 움직이기 위한 최적의 행위를 선택하여 로봇은 움직인다. 이 때 학습한 상황과 달리 움직이는 물체가 있어 학습한 경로로 이동할 수 없는 상황이 발생하는 경우 음수의 보답을 받게되어 Q값을 재학습시킨다. 이 때 움직이는 물체와 만나게 되면 현재 상태에서 선택된 행위의 보답을 음수로 설정하여, Q값을 낮추고 다시 현재 상태에서 최대치를 갖는 행위를 선택하도록 하였다. 최대치를 선택하는데 있어 앞서 사용한 방법과는 달리 상태 s에서 선택한 행위 a의 횟수를 $N(s,a)$라고 정의한 배열에 기록한 후 다음과 같은 확률로 행위를 선택하였다.

$$a = \max_{a \in A} e^{(-N(s,a)*a)} Q(s,a) \qquad (5)$$

(5) 식으로 현재 물체가 가로막고 있는 위치가 절대적으로 우세한 Q값을 가지고 있는 경우 많은 양의 음수의 보답값을 기다리지 않고 우회하는 경로를 선택할 수 있었다.

Fig. 2는 기존에 학습된 값으로 주행을 한 결과이다. 만약 로봇이 움직이는 경로에 물체가 있다면 경로를 수정하여 주어진 목적지로 이동하여야 한다. Fig 3.은 움직이는 물체가 있는 상황에서 모의실험결과이다. 움직이는 물체는 로봇과 목적지사이에 놓여지며 로봇의 움직임에 따라 목적지까지 최단경로의 위치로 이동하여 로봇의 이동을 방해하도록 하였다. 움직이는 물체의 속도와 이동방향은 로봇과 똑같게 하였다. 그림에서 사각형으로 구성된 궤적이 움직이는 물체의 궤적이며, 원형으로 구성된 궤적은 로봇의 궤적이다.

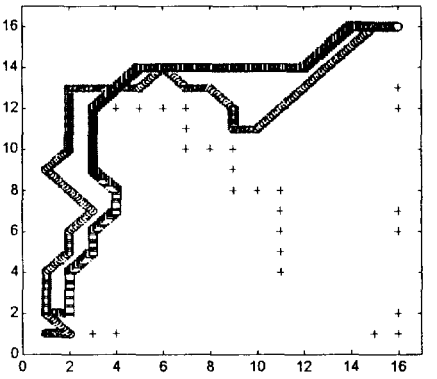

그림 3. 움직이는 물체가 있는 경우의 주행
Fig 3. Path-planning with moving object

만약 움직이는 물체가 로봇이 움직일 수 있는 유일한 경로를 막고 서 있으면 그 상태에 대한 보답은 계속 작아지게 되어 다른 경로를 찾게 된다. Fig. 4에 가운데 부분에 통로를 만들어 놓아 최단경로가 가운데가 되게 학습을 한 후 움직이는 물체를 경로상에 위치하게 하였다. 로봇은 계속적으로 시도하다가 보답값이 작아지면 다른 경로로 주행함을 볼 수 있다.

4.3 여러 대의 로봇이 움직이는 동적환경에서 학습과 방법

여러 대의 로봇이 동일한 공간에서 목표점을 향해 움직일 때 각각의 로봇은 상대방 로봇을 회피하면서 목표점으로 이동하여야 한다. 여러 대의 로봇이 움직이는 경우는 4.2절의 경우를 확장한 것으로 생각할 수 있다. 즉, 자신을 제외한 다른 로봇은 모두 움직이는 물체로 보고 학습을 하면 상대방 로봇을 회피하면서 목적지로 이동하게 된다. 이와 같은 방법으로 여러 대의 로봇이 동일한 환경에서 움직이는 경우에도 목적지까지 모두 이동할 수 있다.

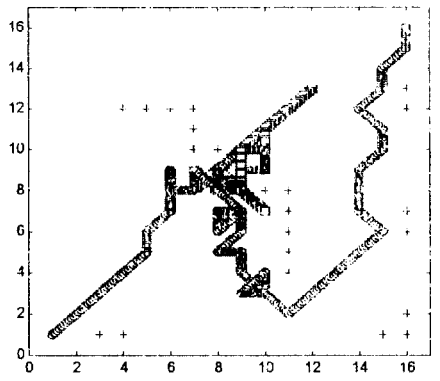

그림 4. 움직이는 물체가 가로막고 있는 경우 다른 경로로 주행
Fig. 4

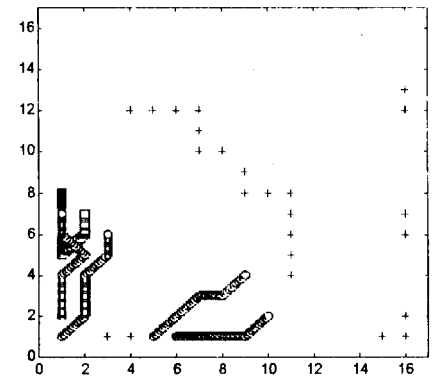

그림 5. 4대의 로봇과 움직이는 물체가 있는 경우
Fig 5. Simulation of robots and obstacle

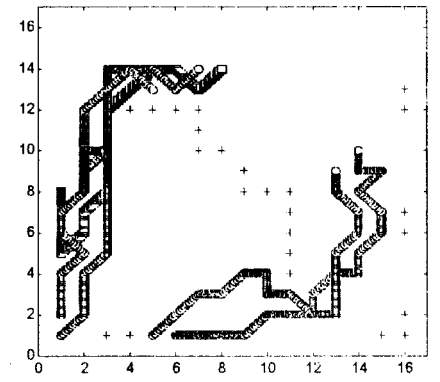

그림 6. 충돌을 피하면서 목적지로 가는 과정
Fig 6. Obstacles and other robots avoidance

5. 결과

미지의 환경에서 적응능력이 우수한 Q-learning을 사용하여 이동로봇의 경로계획과 경로제어를 움직이는 물체가 있는 동적인 환경에 적용하였다. 기존의 방법들은 움직이는 물체의 궤적을 알아야했으나 본 논문에서는 물체의 움직임에 따라 온라인으로 학습을 하여 물체를 회피하도록 하였다. 또한 여러 대의 로봇이 움직이는 경우로 확장하여도 적용이 가능함을 보였다. 이러한 학습 능력을 가지고 있는 로봇은 복도나 공공장소에서 효과적으로 적용이 될 것으로 판단된다.

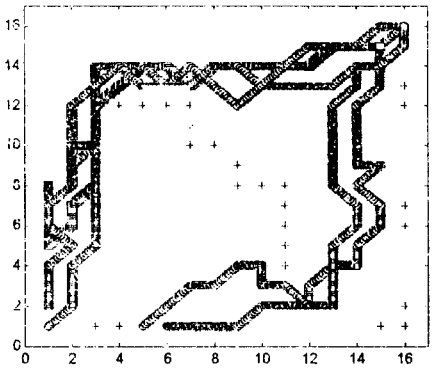

그림 7. 모든 로봇들이 목적지까지 도착한 경로
Fig. 7 All robots succeeded in moving the goal

참고문헌

[1] M. Asada, E.Uchibe, S. Noda, S. Tawaratsumida, and K. Hosoda, "Coordination of Multiple Behaviors Acquired By a Vision-Based Reinforcement Learning", *IROS*, pp.917-924, 1994

[2] R.C. Arkin, "Motor schema-based mobile robot navigation", *International Journal of Robotics Research* Vol 8, pp.92-112, 1989

[3] Brooks R.A. , "A Robust Layered Control System for a Mobile Robot", *IEEE Journal of Robotics and Automation*, Vol RA-3, No 1, March 1986.

[4] O.Causse, L.H. Pampagnim, "Management of a multi-robot system in a public environment", *IROS*, pp.246-252, 1995

[5] L.P. Kaelbling, M.L. Littman and A.W. Moore, "Reinforcement Learning: A Survey", *Journal of Artificial Intelligence Research* pp. 237-285, 1996

[6] M. J. Martaric, M. Nilsson, and K. T. Simsarian, "Cooperative Multi-Robot Box-Pushing",*IROS*, pp,556-561, 1995

[7] S. Russel and P. Norvig, "Artificial Intelligence : A Modern Approach", *Prentice Hall*

[8] T. Tsubouchi, K. Hiraoka, T. Naniwa and S. Arimoto, "A Mobile Robot Navigation Scheme for an Environment with Multiple Moving Obstacles", *IROS*, pp.1791-1798, 1992

[9] Special Issue on Autonomous Intelligent Machines. *IEEE Computer*, Vol22, No.6, June 1989

[10] 서일홍,김재현,오상록, "Fuzzy-Q learning",*제어계측,자동화,로보틱스 연구회 합동학술대회논문집*, pp. 346-350, 1996