Proceedings of KFIS Fall Conference, 2003. 5. 3

강화학습을 이용한 주행경로 최적화 알고리즘 개발

Optimal Route Finding Algorithms based Reinforcement Learning

정희석, 이종수
연세대학교 대학원 기계공학과, 연세대학교 기계공학부

Heeseok Jeong, Jongsoo Lee
Dep. of mechanical engineering, Yonsei Univ., School of mechanical engineering, Yonsei Univ.
E-mail : jleej@yonsei.ac.kr

ABSTRACT

본 논문에서는 차량의 주행경로 최적화를 위해 강화학습 개념을 적용하고자 한다. 강화학습의 특징은 관심 대상에 대한 구체적인 지배 규칙의 정보 없이도 최적화된 행동 방식을 학습시킬 수 있는 특징이 있어서, 실제 차량의 주행경로와 같이 여러 교통정보 및 시간에 따른 변화 등에 대한 복잡한 고려가 필요한 시스템에 적합하다. 또한 학습을 위한 강화(보상, 벌칙)의 정도 및 기준을 조절해 줌으로써 다양한 최적주행경로를 제공할 수 있다. 따라서, 본 논문에서는 강화학습 알고리즘을 이용하여 다양한 최적주행경로를 제공해 주는 시스템을 구현한다.

Key words : 강화학습(Reinforcement Learning), Q-learning, 최적주행경로,

I. 서 론

최적경로의 탐색은 주어진 네트워크에서 임의의 시작점과 종점 사이를 연결하는 경로들 중 원하는 목적을 가장 효과적으로 달성할 수 있는 경로를 탐색하는 문제라고 정의할 수 있다. 전통적으로 이러한 최적경로 탐색 알고리즘은 운송분야 등에서 운송비용 및 시간의 최적화 문제에 많이 적용되어 왔다. 여기에 최근 들어서는 통신 및 물류분야 등에서 각 경로에 할당되는 부하(load)들을 조절하여 전체적인 네트워크를 최적화하는 여러 알고리즘 들이 제안되고 또 실제로 적용되고 있다.

특히, 교통량 증가에도 불구하고 신규 도로 건설이 어렵기 때문에 기존 도로의 사용효율을 극대화하기 위한 교통량 분산을 위해 경로 최적화 알고리즘의 적용 방법에 대한 연구가 진행되고 있다.

그러나, 도로의 특성상 기존의 경로 최적화 방법을 그대로 적용하기 곤란한 점이 있다. 회전제약과 교통상황의 시간에 따른 변화를 그 예라 할 수 있다. 따라서, 전통적인 최적경로 탐색 알고리즘에 이러한 교통 특성을 부여하기 위해 기존 알고리즘의 수정 또는 새로운 알고리즘의 개발 및 연구가 수행되고 있다.

본 논문은 최적주행경로 탐색을 위해 강화학습 개념을 적용하고자 한다. 강화학습은 목적을 달성할 수 있는 행동에 보상을 부여함으로써 최적의 행동 집합을 찾아내는 알고리즘이다. 특히, 실제의 교통상황 등 복잡한 요인으로 인해 통행비용이 변할 때 이를 해석하기 위한 정확한 지배방정식이 필요치 않다는 장점이 있다. 또한 학습을 위한 보상조건을 적절히 조절해 줌으로써 다양한 최적주행경로(최단거리 경로, 최단시간 경로, 운전 용이 경로 등) 탐색이 가능하다. 결국, 본 논문에서는 회전제약이 고려

되고 다양한 최적경로를 제공할 수 있는 강화학습 기반의 주행경로 최적화 알고리즘을 제안하고자 한다.

II. 본 론

2.1 강화학습과 Q-learning

강화학습[1, 2]은 작동 환경에 대한 학습 모형을 사용하지 않는 대표적인 기계학습방법 중 하나이다. 강화학습의 학습대리자(agent)는 현재 상태에서 실행된 행동에 의해 기대함수 값을 계산하고 각 상태에서 가장 적절한 행동을 선택하는 기준(정책, policy)을 만들어낸다. 특히, 바로 이전 단계의 행동을 제외하고는 지난 과거의 이력(history)을 기억하지 않는다는 특징이 있다.

Q-learning은 강화학습 방법의 하나로 행동가치함수(action-value function)를 이용해서 학습을 수행한다. 가치함수는 다음의 (1)로 정의 된다.

$$V(s) = \max_a Q(s,a) \qquad (1)$$

여기서, s는 임의의 상태를, a는 선택된 행동을 나타낸다. 즉, 임의의 상태 s에서의 가치함수는 그 상태에서 Q값이 최대로 되는 행동에 의해 계산된 Q값으로 정의된다. 따라서, 목표를 달성하기 위한 최적의 정책은 각 상태에서 Q값을 최대로 하는 행동들의 집합으로 정의할 수 있다.

Q-learning의 일반적인 학습과정은 임의의 상태에서 행동을 결정하고 이에 따른 보상을 이용하여 Q값을 갱신하는 과정을 반복하는 방법으로 학습이 수행된다. 이때, Q값의 갱신은 일반적으로 (2)의 식을 따른다.

$$Q(s,a) \leftarrow (1-\alpha) * Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a')] \qquad (2)$$

여기서, r은 보상, α는 step size, γ는 감소율을 의미한다.

본 논문에서는 학습 과정에서 다음 상태로의 진행을 위해 선택되는 행동은 ε-greedy 방법을 이용하여 결정하였다. ε-greedy 방법은 행동의 선택을 위해 기본적으로 현재 구성된 Q값들 중 가장 큰 값을 갖는 행동을 선택한다. 그러나, ε ($0 < \varepsilon <= 1$)의 확률만큼은 임의의 행동을 취하도록 해서 다양한 행동 집합에 의해 학습이 진행되도록 하는 선택 방법이다. 일반적으로 이 경우 ε은 선택이 진행되어 상태가 변할수록 작아지도록 조절하여 일정한 선택이 진행 된 상태에서는 다양한 행동이 선택되기보다는 목표에 도달하는 행동이 선택될 확률이 높아지도록 해 주는 것이 일반적이다.

2.2 주행경로의 특성

2.2.1 회전제약

전통적인 경로최적화 알고리즘에서는 직진, 좌회전, 우회전, U턴이 모두 허용된다고 가정한다. 그러나, 실제의 도로에서는 도로구조, 교통량 및 도로크기에 따라, 진행방향 중 일부를 제한하고 있다. 예를 들어, 오거리의 경우 교차로에 진입하는 차량에 대해 좌회전을 허용할 경우 통행 자체가 불가능하게 된다. 따라서, 기존의 알고리즘에 회전제약을 표현할 수 있도록 전노드나 전전노드 등의 정보를 포함하는 알고리즘이 제안되기도 하였다. 또한, 임의 경로를 생성하고 이중 더 우수한 경로(단거리, 단시간 등)의 특성을 반영하기위해 유전자알고리즘을 적용하는 연구도 수행되었다[3]. 이 경우 경로생성시에 회전제약을 고려하도록 제안되었다.

본 논문에서는 각 도로의 회전제약을 고려하도록 수정된 학습구조 및 교차로 정보구조를 제안한다.

2.2.2 시간에 따른 변화

기존의 경로 최적화 알고리즘의 경우 각 도로의 통행비용은 시간에 대해 일정한 것으로 가정한다. 이는 차량의 출발에서 도착까지 운행시간이 통행량 변화시간에 비해 충분히 짧다는 가정이 포함되어 있다. 그러나, 시내 교통상황의 경우 비교적 짧은 시간에 교통량이 늘어나고, 정체상황이 빠르게 인근지역으로 전파되는 특징이 있다.

결국, 기존의 알고리즘을 동적인 네트워크로 확장하기 위해서는 통행비용 함수의 시간변화를 고려하여 복잡한 탐색을 수행해야 한다. 그러나, 본 논문에서 제안하는 알고리즘은 각 교차로에서의 통행비용만을 고려하기 때문에 임의의 시간에 대한 통행비용 함수만 제공된다면 시간에 따른 통행비용 변화를 고려한 최적경로 탐색 알고리즘으로의 확장이 가능하다.

2.3 주행경로 최적화 알고리즘

2.3.1 교차로 및 도로의 정보구조

본 논문에서는 앞서 고려한 회전제약을 표현하기 위해 각 교차로의 정보구조에 통행가능 경로를 포함시켰다. 다시 말해, 그림 1과 같이 2번 도로의 경우 좌회전이 금지되어 있기 때문에 주어진 교차로에서 실제 가능한 경로는 5개이고 각각의 경로를 진입-진출도로의 집합으로 표시하여 정보구조에 포함시켰다.

도로 정보의 경우 양단에 위치한 교차로 정

보를 이용하여 표시하였고 실제 도로의 경우 곡선주로 등의 특성을 반영하기 위해 실제 거리를 정보구조에 포함하였다.

2.3.2 보상규칙

본 논문에서 사용한 보상규칙은 크게 두 가지이다. 먼저, 목적지 도달에 관한 보상이다. 이를 위해서, 현재 교차로에서 목적지 교차로까지의 직선거리를 보상의 기준으로 삼았다. 즉, 출발 교차로와 목적지 교차로까지의 직선거리에 대한 현재/다음교차로와 목적지까지의 직선거리의 비를 보상 기준으로 사용하였고, 이는 (3)의 식으로 표현된다.

$$r_1 = r_{base} * (d_{present} - d_{next}) / d_0 \quad (3)$$

여기서 r_{base}는 기준 보상, d_0는 출발교차로에서 목적지 교차로까지의 직선거리, $d_{present}$는 현재 교차로에서 목적지까지의 직선거리, d_{next}는 다음 교차로에서 목적지까지의 직선거리를 나타낸다.

두 번째 보상규칙은 진행한 거리에 대한 벌칙에 의해 계산된다. 이는 다음 교차로에서 목적지까지의 직선거리가 동일한 경우 현재 교차로에서 가까운 교차로로 진행하도록 하기 위해서 사용되며 (4)의 식으로 표현된다.

$$r_2 = r_{base} * (-1) * d_{rel} / d_0 \quad (4)$$

여기서 d_{rel}은 두 교차로 사이의 거리이다. 실제 주행거리를 적용하기 위해 각 도로의 정보구조에 포함되어 있는 실제 길이를 이용한다.

전체 보상은 (3), (4)에서 구한 r_1, r_2의 합으로 구성되며, 각 교차로에서 적은 거리를 주행하면서 목적지에 도달할 수 있는 경로가 선택되도록 보상규칙을 정의하였다.

이외에도 교차로 통과 횟수 등의 여러 가지 요인들에 대한 보상규칙 설정을 통해 다양한 최적주행 경로를 탐색하는 것이 가능하다.

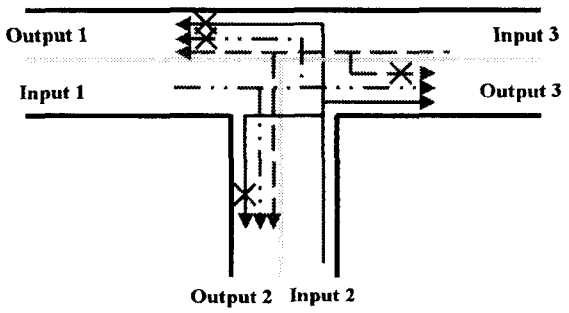

그림 1. 통과가능 경로를 포함하는 교차로 정보

2.3.3 회전제약을 고려한 Q-learning

본 논문에서 제안하는 Q-learning에서는 진행할 다음 교차로를 선택할 때 현재 교차로에서 진출이 가능한 지를 먼저 확인하도록 하였다. 이는 회전제약(좌회전/U턴 금지 등)이 적용되는 경로들을 선택대상에서 제외되도록 하였다. 이를 통해 각 교차로에서 회전제약이 포함된 경로가 원천적으로 생성되지 않도록 하였다.

2.3.4 주행경로 최적화 알고리즘의 구성

그림 2는 본 논문에서 제안하는 최단거리 주행경로 제공 알고리즘의 흐름도이다.

먼저, 관심지역의 교차로 및 도로 정보를 확인하고 임의의 교차로에서 진행방향을 결정하고 2.3.2에서 정의한 보상규칙에 의해 보상을 계산하고 (2)식을 이용하여 각 교차로에서의 경로에 대한 Q값을 갱신한다. 다음 교차로로 선택된 것이 목적지 교차로일 때까지 진행방향 선택-보상의 과정을 반복하고 목적지에 도달하면, 출발-선택-보상-도착의 과정을 주어진 횟수만큼 반복하여 학습을 수행한다.

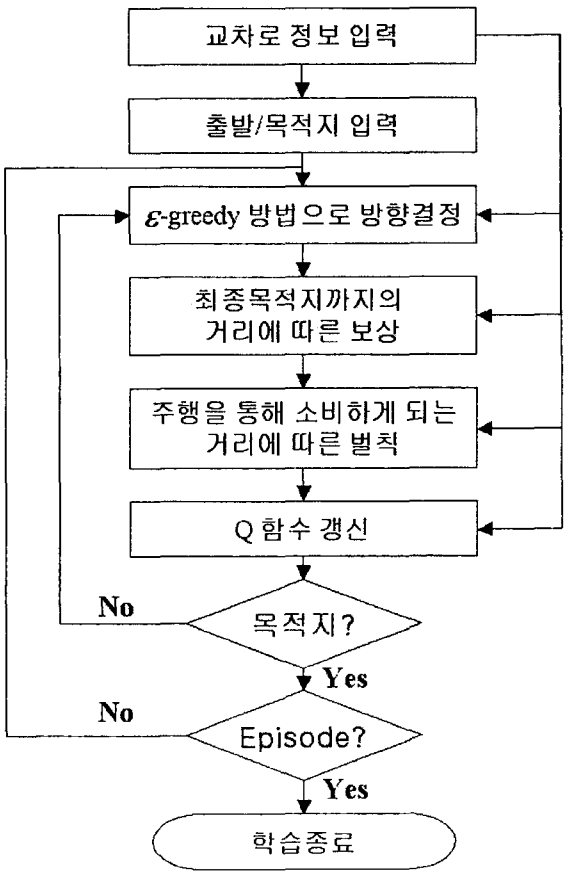

그림 2. 제안하는 최적주행경로 탐색 알고리즘

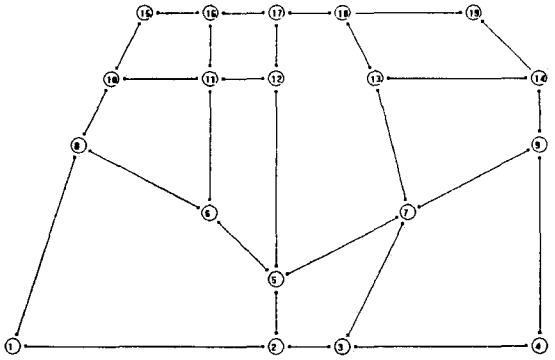

그림 3. 검증용 교통네트워크

표1. 네트워크 구성 교차로 및 거리 정보

번호	위치		교차로까지 거리(교차로 번호)			
	x	y				
1	0	0	40(2)	31.62(8)		
2	40	0	40(1)	10(3)	10(5)	
3	50	0	10(2)	30(4)	22.36(7)	
4	80	0	30(3)	30(9)		
5	40	10	10(2)	14.14(6)	22.36(7)	30(12)
6	30	20	14.14(5)	22.36(8)	20(11)	
7	60	20	22.36(3)	22.36(5)	22.36(9)	20.62(13)
8	10	30	31.62(1)	22.36(6)	11.18(10)	
9	80	30	30(4)	22.36(7)	10(14)	
10	15	40	11.18(8)	15(11)	11.18(15)	
11	30	40	20(6)	15(10)	10(12)	10(16)
12	40	40	30(5)	10(11)	10(17)	
13	55	40	20.62(7)	25(14)	11.18(18)	
14	80	40	10(9)	25(13)	14.14(19)	
15	20	50	11.18(10)	10(16)		
16	30	50	10(11)	10(15)	10(17)	
17	40	50	10(12)	10(16)	10(18)	
18	50	50	11.18(13)	10(17)	20(19)	
19	70	50	14.14(14)	20(18)		

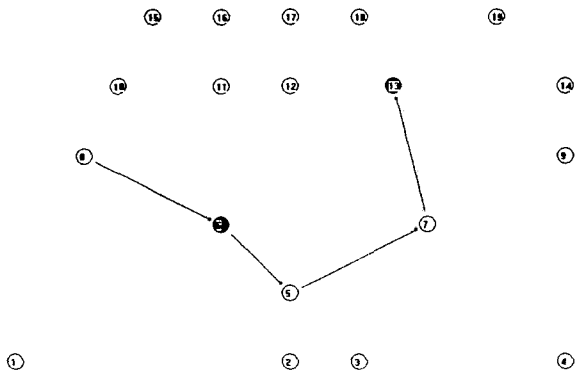

그림 4. 최단거리 주행경로 탐색 결과
(8번 교차로->6번 교차로 방향으로 출발, 13번 교차로 도착을 위한 최단거리 주행경로 : 8->6->5->7>13)

2.4 최단거리 주행경로의 탐색

그림 3은 제안된 알고리즘을 검증하기 위해 구성한 교통 네트워크이고, 표1은 각 교차로의 2차원 위치와 인근 교차로까지의 직선거리를 표시한 것이다. 여기서, 각 도로의 길이는 교차로 사이의 직선거리로 정의하였다. 또한, 각 교차로에서의 회전제약은 무시하였다.

교차로 8번에서 6번 교차로 방향으로 출발하여 목적지 교차로 13번에 도착하는 가능한 최단거리 주행경로는 표에서 계산하면 아래와 같다.

경로1 : 8->6->11->12->17->18->13
(8->6->11->16->17->18->13)
경로2 : 8->6->5->12->17->18->13
경로3 : 8->6->5->7->13

경로1의 주행거리 :
22.36+20+10+10+10+11.18=83.54
경로2의 주행거리 :
22.36+14.14+30+10+10+11.18=97.68
경로3의 주행거리 :
22.36+14.14+22.36+20.62=79.48

경로3이 최단거리 주행경로임을 확인할 수 있다. 이에 대해 제안된 알고리즘에 의해 탐색된 최단거리 주행경로는 그림 4에서와 같이 표에 의한 직접 계산의 결과와 동일하다.

III. 결 론

본 논문에서는 회전제약 및 시간변화에 따른 통행비용을 고려한 강화학습 기반의 최적주행경로 탐색 알고리즘을 제안하였다. 또한, 간단한 교통네트워크를 적용하여 회전제약이 없는 정적 네트워크에서의 타당성을 검증하였다.

감사의 글 : 본 연구는 하나로 통신㈜의 연구 비지원으로 수행되었습니다.

IV. 참고문헌

[1] Richard S. Sutton and Andrew G. Barto, "Reinforcement Learning", The MIT Press, 2002
[2] Tom M. Mitchell, "Machine Learning", The McGraw-Hill Companies, Inc., 1997, pp. 367-390
[3] 김성수, 정종두, 이종현, "유전자알고리즘을 사

용하여 다수최적경로를 제공할 수 있는 동적경로유도시스템의 개발", IE Interfaces, Vol. 14, No. 4, December 2001, pp. 374-384