마르코프 결정 프로세스의 위상적 계산 복잡도 척도

이승준° 장병탁
서울대학교 바이오지능연구실
sjlee°@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

Topological measures for algorithm complexity of Markov decision processes

Seung-joon Yi° Byoung-Tak Zhang
SNU Biointelligence lab

요 약

실세계의 여러 문제들은 마르코프 결정 문제(Markov decision problem, MDP)로 표현될 수 있고, 이 MDP는 모델이 알려진 경우에는 평가치 반복(value iteration) 이나 모델이 알려지지 않은 경우에도 강화 학습(reinforcement learning) 알고리즘 등을 사용하여 풀 수 있다. 하지만 이들 알고리즘들은 시간 복잡도가 높아 크기가 큰 실세계 문제에 적용하기 쉽지 않아, MDP를 계층적으로 분할하거나, 여러 단계를 묶어서 수행하는 등의 시간적 추상화(temporal abstraction) 방법이 제안되어 왔다.

이러한 시간적 추상화 방법들의 문제점으로는 시간적 추상화의 디자인에 따라 MDP의 풀이 성능이 크게 달라질 수 있으며, 많은 경우 사용자가 이 디자인을 직접 제공해야 한다는 것들이 있다. 최근 사용자의 간섭이 필요 없이 자동적으로 시간적 추상화를 만드는 방법들이 제안된 바 있으나, 이들 방법들 역시 결과물에 대한 이론적인 성능 보장(performance guarantee)은 제공하지 못하고 있다.

본 연구에서는 이러한 문제점을 해결하기 위해 MDP의 구조와 그 풀이 성능을 연관짓는 복잡도 척도에 대해 살펴본다. 이를 위해 MDP로부터 얻은 상태 경로 그래프(state trajectory graph)의 위상적 성질들을 여러 네트워크 척도(network measurements) 들을 이용하여 측정하고, 이와 MDP의 풀이 성능과의 관계를 다양한 상황에 대해 실험적, 이론적으로 분석해 보았다.

1. 서 론

실세계의 다양한 문제들은 상태(state) 와 상태에서 취할 수 있는 행동(action), 그리고 상태와 행동의 결과인 보상(reward)의 형태로 주어지는 마르코프 결정 문제(Markov decision problem, MDP)의 형태로 표현될 수 있다. MDP를 푼다는 것은 최적의 보상을 얻는 정책(policy)를 구하는 것을 의미하며, MDP는 환경의 모델을 알 경우 동적 프로그래밍(dynamic programming)등을 이용하여 풀 수 있고, 환경이 알려져 있지 않은 경우에도 시행착오를 통해 학습하는 강화 학습 (reinforcement learning)을 사용하여 풀 수 있다.

이러한 방법들은 계산 복잡도가 높고, 기본적으로 이산적인 시간과 공간을 가정하기 때문에 연속적인 시간과 공간을 갖는 실세계 문제에 바로 적용하기가 어렵다. 주로 쓰이는 대안으로는 신경망과 같은 함수 근사장치를 사용하는 방법으로 많은 성공적인 적용 예가 있어 왔으나, 이러한 근사 장치를 사용하더라도 문제가 복잡해짐에 따라 학습해야 할 파라미터의 수가 증가하는 것은 막을 수가 없다. 강화 학습의 경우 학습 난이도가 문제 크기나 파라미터 수가 커짐에 따라 지수적으로 증가하는 차원성의 저주[1] (curse of dimensionality) 문제 때문에 크고 복잡한 실세계 문제에 적용하기가 매우 어렵다.

이러한 문제를 해결하기 위해 제안된 방법이 시간적 추상화(temporal abstraction) 방법이다. MDP가 크기가 커짐에 따라 두 상태 간에 더 많은 판단 단계를 거치게 되어 학습이 어려워지는 것을 막기 위해, MDP를 계층적으로 분할하거나[2] 한 번에 실행되는 여러 행동들의 묶음인 option을 사용한다[3]. 이러한 방법으로 요구되는 판단 단계의 수를 줄여, 크기가 큰 MDP를 보다 효율적으로 풀 수 있게 해 준다. 이들 방법들은 실제 문제에 적용 시 MDP의 풀이 속도를 크게 향상시킴이 보여 왔고, 기존 방법으로 불가능한 크기가 큰 MDP도 풀 수 있음이 보여 왔다[1].

반면 시간적 추상화 방법들의 문제점으로는 우선 대부분의 경우 사용자가 문제에 대한 지식을 바탕으로 직접 시간적 추상화의 디자인을 해야 하고, 또한 이러한 시간적 추상화의 디자인에 따라 MDP의 풀이 성능이 크게 달라질 수 있다는 것이다[4]. 첫 번째 문제점을 해결하기 위해 최근에는 자동적으로 시간적 추상화 디자인을 하는 여러 가지 방법들이 제안되어 오고 있으나[5,6,7,8], 이러한 방법들 역시 MDP의 풀이 성능에 대한 성능 보장(performance guarantee) 는 하지 못하고 있다.

시간적 추상화 디자인 시에 MDP의 풀이 성능을 보장하기 위해서는 우선 MDP의 시간적 추상화 디자인으로부터 MDP의 풀이 성능을 평가할 수 있는 척도(measurement) 가 필요하다. 이러한 척도가 존재한다면 이를 이용하여 우수한 풀이 성능을 갖도록 시간적 추상화 디자인이 가능하고, 더 나아가 MDP의 풀이 성능에 대한 성능 보장 또한 가능해지게 되나, 이러한 성능 척도를 명시적으로 사용한 방법은 기존에 찾아볼 수 없었다.

본 논문에서는 이러한 척도를 얻기 위해 MDP로부터 얻어지는 상태 경로 그래프(State trajectory graph)의 위상학적 성질들을 여러 네트워크 척도(Network measurements)들을 사용하여 분석하고, MDP의 풀이 성능과의 관계를 다양한 상황에 대해 실험적으로 분석하고 이론적인 성능 보장 결과를 얻었다.

2. 관련 연구

2.1. MDP와 시간적 추상화(temporal abstraction)

MDP는 $<S,A,T,R>$로 정의할 수 있다[9]. S는 유한한 상태들 s의 집합, A는 유한한 행동들 a의 집합, R은 (s,a,s') 의 함수로 주어지는 보상, T 는 행동 a를 수행할 경우 상태 s에서 s'로 이동할 확률이다. 정책 π는 각 상태에서 취할 행동들로, MDP를 푸는 것은 장기적으로 최고의 보상을 얻을 수 있는 최적 정책 π*를 구하는 것이다. MDP의 모델, 즉 R,T에 대해 알고 있을 경우 식 (1)의 평가치 반복(value iteration) 등의 동적 프로그래밍 방법으로 π*를 구할 수 있고, 모델에 대해 모를 경우에도 식 (2)의 Q-학습(Q-learning) 등의 강화학습 알고리즘을 사용하여 π*를 구할 수 있다.

$$V(s) \leftarrow V(s) + \alpha[r + \gamma \max_{s'} V(s') - V(s)] \quad (1)$$

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

MDP를 풀 때의 시간적 복잡도를 낮추기 위해 제안된 방법이 여러 단계의 행동을 한꺼번에 수행할 수 있게 하는 시간적 추상화(temporal abstraction) 방법들이다[1]. MDP에서는 각 행동들에 단위 시간이 소요된다는 것을 가정하나, 각 행동들에 다양한 시간이 소요될 수 있도록 MDP 모델을 확장한 Semi MDP를 사용하여 이러한 시간적 추상화를 나타낼 수 있다. 이 경우 상태 s에서 확장된 행동 o를 수행할 경우 소요되는 시간을 $t(s,o)$로 나타낸다. Semi MDP의 경우에도, 평가치 반복이나 Q-학습 등의 방법을 아래의 식 (3),(4) 와 같이 약간의 수정을 거쳐 적용 가능하다.

$$V(s) \leftarrow V(s) + \alpha[r + \max_{s',o} \gamma^{t(s,o)} V(s') - V(s)] \quad (3)$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \max_{o} \gamma^{t(s,o)} Q(s,o') - V(s)] \quad (4)$$

이러한 시간적 추상화 방법들의 가장 큰 단점은 사용자가 직접 계층 구조를 디자인해야 하는 것으로, 이를 해결하기 위해 시간적 추상화를 자동화하려는 다양한 시도들이 있어 왔다. 현재까지의 시도들을 크게 분류해 보면 상태의 방문 빈도를 사용하여 중요한 상태를 찾아내는 방법[5], 특정 조건을 사용해 상태들을 반복적으로 분할하는 방법[6], 특정 조건을 만족하는 중요한 상태를 찾아내는 방법[7], 마지막으로 MDP로부터 대응되는 그래프를 얻고 여기에 그래프 이론 기반 알고리즘을 사용하는 방법[8] 등으로 나눌 수 있다. 본 연구에서는 그래프 이론 기반 방법들의 접근 방식을 채용하고 있다. 이에 대해서는 아래 장에서 자세히 서술하기로 한다.

2.2. 상태 경로 그래프(State Trajectory Graph)

위에서 서술한 바와 같이, 그래프 이론 기반 알고리즘을 MDP의 성능 향상에 이용하기 위해서는 MDP로부터 이에 대응되는 상태 경로 그래프를 얻는 방법이 있다.

상태 경로 그래프란 MDP의 각 상태 s_i를 그래프의 각 노드 i에 대응시키고, MDP의 상태 변화 (s_i,a,s_j)들을 각 에지 (i,j)에 대응시켜 생성한 그래프이다. 결정론적인 MDP의 경우 각 행동의 결과로 일어나는 상태 변화는 유일하므로 네트워크의 각 에지 (i,j)는 MDP의 상태와 행동 (s,a)에 대응하게 된다. 이동 확률(transition probability) T를 갖는 확률적인 MDP의 경우 네트워크의 각 에지에는 상태, 행동 (s,a)의 조합으로부터 상태 변화 (s_i,a,s_j)가 일어날 확률 $T_{i,j}$가 가중치(weight) 로 주어지는 형태로 나타낼 수 있다.

이와 같이 주어진 MDP를 상태 경로 그래프의 형태로 나타낼 수 있으며, 이러한 방법을 통해 MDP에 그래프 이론 기반 알고리즘을 적용할 수 있다[8]. 본 연구에서도 이와 같은 방법을 사용하였다.

3. MDP의 성능과 위상적 척도

3.1. MDP의 위상적 성질과 풀이 성능

전술했듯이 MDP의 풀이 성능을 향상시키기 위해 제안된 것이 시간적 추상화 방법들이고, 이를 자동화하기 위해 여러 자동적인 시간적 추상화 방법들이 제안되어 왔다. 허나 이러한 방법들의 공통적인 문제점은 얻어지는 MDP의 성능에 대한 보장이 없다는 것이다. 구체적인 예로 다음과 같은 세 모델을 비교해 볼 수 있다.

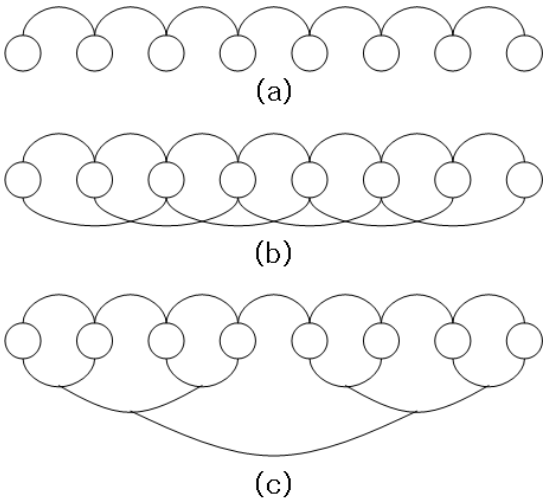

그림 1. 세 종류의 MDP 구조

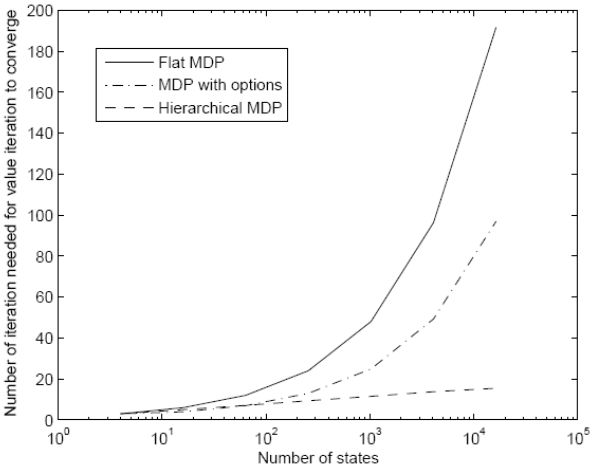

그림 2. MDP의 종류에 따른 사이즈와 풀이 시간 사이의 관계

(a)의 경우 각 상태들이 인접한 상태들과 연결되어 있는 MDP이고, (b)는 여기에 size 2인 option[3]을 균일하게 추가한 경우, (c)의 경우 size 2인 상태들을 계층적으로 모아 상위 MDP를 형성해 감으로서 계층적인 구조를 형성한 경우[2]이다. 위와 같은 형태를 가진 MDP에 대해 한 state에서만 보상 1을 책정하고 평가치 반복 알고리즘을 적용하여 풀 경우, MDP의 사이즈에 따른 총 풀이 시간과의 관계는 그림 2와 같은 양상을 보인다.

(a)의 경우 문제 사이즈가 커짐에 따라 그에 비례하여 시간이 걸리게 되고, option을 추가한 (b)의 경우 시간이 줄어들지만 역시 문제 사이즈에 비례하는 모습을 보인다. 반면 (c)의 경우 비슷한 수의 option을 추가했음에도 풀이 시간이 문제 사이즈의 로그값에 비례하는 모습을 볼 수 있다. 즉, 시간적 추상화를 적용할 경우 MDP의 풀이 성능은 향상되지만, 시간적 추상화의 위상적 구조에 따라 문제 사이즈에 따른 성능의 관계가 크게 달라질 수 있기에, MDP의 구조와 풀이 성능간의 관계를 규정짓는 척도가 필요하다는 것을 알 수 있다.

3.2. 위상적 척도(Topological measurements)

최근의 연구 결과 실세계의 많은 네트워크들은 작은 세상 성질 (small world property) 나 높은 클러스터링 계수[10] (clustering coefficient), 척도 없는 도수 분포[11](scale-free degree distribution) 등의 여러 특징적인 성질을 공유하는 것이 밝혀졌다. 이러한 네트워크의 위상적 성질들은 네트워크의 연결 관계를 특징짓고 그 네트워크상에서 일어나는 여러 과정들에 영향을 준다. 이들을 측정하는 데 사용되는 것이 네트워크 척도(network measurement)로써 네트워크들을 분석하고 서로 다른 종류들로 구분하며, 원하는 성질을 갖는 네트워크를 디자인하는 등 다양한 분야에 사용된다[12].

이러한 위상적 성질들과 MDP 풀이 효율간의 관계는 기존 연구[13]에서 실험적으로 분석된 바 있다. Regular lattice 모델, Erdos-renyl 모델, Watts-Strogatz 모델, Barabasi-Albert 모델들을 사용하여 여러 크기와 생성

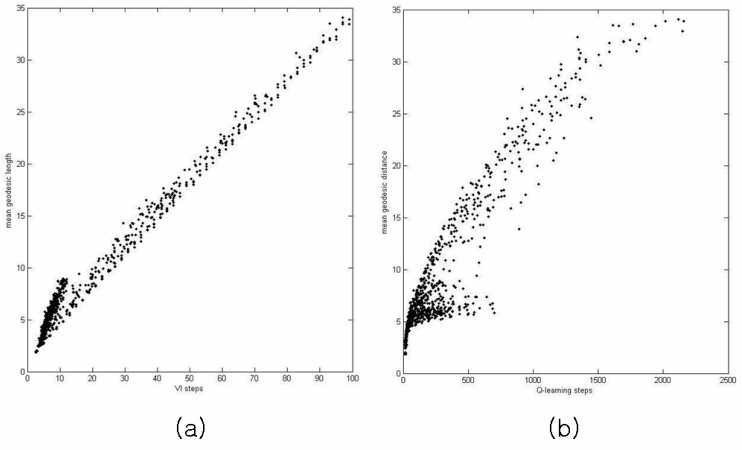

그림 3. 평균 측지 거리와 MDP의 풀이 성능간의 상관 관계. (a): 평가치 반복 (b): Q-Learning

파라미터를 사용한 다양한 MDP를 생성하였고, 이 MDP를 평가치 반복과 Q-학습을 사용하여 풀이 시간을 각각 측정하고 MDP의 상태 변화 그래프로부터 평균 차수, 최대 차수, 클러스터링 계수 , 부분 그래프 중앙도, 프랙탈 차원 등의 다양한 네트워크 척도를 측정하였다.

실험 결과 타 네트워크 척도들과 달리, 평균 측지 거리(Mean geodesic distance)는 그림 3과 같이 평가치 반복과 Q학습 모두에 대해 다양한 네트워크 모델과 파라미터를 사용하더라도 공통적으로 관측되는 뚜렷한 상관관계가 있어, MDP의 성능을 가리키는 척도로 사용할 수 있음을 알 수 있었다.

3.3. 이론적 성능 보장 (Theoretical guarantee)

3.3.1. Deterministic single-reward MDP

가장 간단한 경우로, 상태 s에서 행동 a를 취할 경우 고정된 보상 r과 고정된 다음 상태 s'를 얻는 결정론적 MDP에 대해 알아보자. 또한 이러한 결정론적 MDP 중에서 다시 양의 보상이 한 상태에서만 주어지는 단일 보상 MDP(Single-goal MDP)에 대해 먼저 알아보기로 한다.

우선 평가치 반복의 경우, 식(1) 의 업데이트 룰에 의하여 각 상태는 매 단계마다 주위의 모든 상태들에 대해 backup을 하게 된다[9]. 따라서 특정 상태 s 의 평가함수 $V(s)$는 보상이 주어지는 상태 s_{goal}로부터의 측지거리 $GD(s,s_{goal})$만큼의 단계가 지난 후에 갱신되게 되며, 양의 보상이 한 상태에서만 주어지기 때문에 이 $V(s)$값은 더 이상 변경되지 않는다. 따라서 $V(s)$값은 $GD(s,s_{goal})$ 만큼의 시간 만에 구해지게 되며, 평균적인 풀이 시간은

$$\frac{1}{|s|}\frac{1}{|s|}\sum_{s_1}\sum_{s_2} GD(s_1,s_2) = MGD \qquad (5)$$

즉 단일보상의 결정론적 MDP의 경우 평가치 반복 알고리즘의 기대 수행 시간은 MDP의 상태 변화 그래프의 평균 측지 거리와 일치하게 된다.

Q-학습의 경우 식(2)의 업데이트 룰에 의해 Off-policy backup을 행한다[9]. 따라서 행동을 선택하는 정책에

따라 결과가 달라질 수 있다. 대표적으로 사용하는 정책인 ϵ-greedy의 경우, 확률 ϵ에 따라 임의 탐색(Random search)를 행하고, 나머지의 경우 현재까지 발견된 최적인 행동을 취하는 정책이다. 이 경우, 상태 경로 그래프 상의 최대 차수를 $d_{\max}$라고 하면 각 상태 s가 주위 모든 상태를 backup하는데 걸리는 기대 시간은 $\frac{d_{\max}}{\epsilon}$의 low bound를 갖게 된다. 이 경우 특정 상태 s의 평가 함수 $V(s)$는 $\frac{d_{\max}}{\epsilon} GD(s,s_{goal})$만큼의 기대 시간 만에 구해지게 되며, 평균적인 풀이 시간은 $\frac{d_{\max}}{\epsilon} MGD$의 low bound를 가지게 된다. 따라서 평가치 반복의 경우 풀이 시간은 상태 변화 그래프의 평균 측지 거리에 선형적으로 의존(Linearly dependent)하며, Q-학습의 경우에도 작은 $d_{\max}$를 가지는 MDP의 경우 풀이 시간의 기대치가 MDP의 상태 변화 그래프의 평균 측지 거리에 bounded된다는 것을 알 수 있다.

3.3.2. Stochastic single-reward MDP

보다 일반적인 비결정론적 MDP (stochastic MDP)는 상태 s에서 행동 a를 취할 경우의 결과가 고정되어 있지 않고, 상태 변화 함수 $T(s,a,s')$와 보상 함수 $r(s,a,s')$에 의해 주어지는 MDP이다. 이러한 경우의 한 극단적인 예로는 $T(\bullet,\bullet,s)$의 모든 수치가 0보다 큰 경우가 되겠는데, 이 경우에는 상태 변화 그래프가 완전 그래프(Complete graph)가 되어 MDP의 위상 구조를 따지는 의미가 사라지게 된다. 따라서 본 연구에서는 보다 제한적인 경우로, 상태 s의 다음 상태 s'의 경우의 수가 문제의 크기보다 작은 상한 $d_{\max}$를 갖는 Sparse한 비결정론적 MDP에 대해서만 논의를 전개하도록 한다.

이 경우 $T(\bullet,\bullet,s)$들 중 0이 아닌 최소값을 $t_{\min}$이라 하자. 상태 경로 그래프 상에서 최대 차수는 $d_{\max}$가 되고, 특정 상태 s로부터 가능한 모든 주위 상태들로 backup하는데 걸리는 기대 시간은 위와 같은 과정을 통하면 평가치 반복의 경우 $\frac{1}{t_{\min}}$, Q-학습의 경우 $\frac{d_{\max}}{\epsilon t_{\min}}$이 되게 된다. 따라서, Sparse한 비결정론적 MDP의 경우 평가치 반복의 수행 기대시간은 $\frac{MGD}{t_{\min}}$, Q-학습의 수행 기대시간은 $\frac{d_{\max} MGD}{\epsilon t_{\min}}$ 의 low bound를 갖는다. 따라서, 비결정론적 MDP의 경우에도 특정 조건 하에서는 풀이 시간의 기대치가 역시 MDP의 상태 변화 그래프의 평균 측지 거리에 bounded된다는 것을 알 수 있다.

3.3.3. Multiple-reward MDP

앞서 다룬 단일 보상 MDP보다 일반적인 형태는 양의 보상이 여러 군데에서 주어질 수 있는 다중 보상 MDP(Multi-reward MDP)이다. 이 경우의 문제는 상태 경로 그래프 상에 보상이 0이 아닌 Loop가 존재할 경우 상태 함수가 유한한 단계 안에 수렴하지 않을 수 있다는 것이다.

보다 한정적인 경우로 보상이 0 이상인 Loop가 존재하지 않고 보상이 주어지는 상태들 $s_{r_1}, s_{r_2}, \ldots s_{r_{\max}}$의 수가 총 상태의 수보다 작은 상한 $r_{\max}$를 갖는 경우, 즉 보상이 Sparse한 경우 특정 상태 s의 평가 함수 $V(s)$는 결정론적 MDP의 경우 $\max_i GD(s,s_{r_i})$의 시간 만에 구해지게 되고, 이는 다시 상태 변화 그래프의 최대 측지 거리에 bounded되게 된다. Q-학습의 경우와 위에서 가정한 조건들을 만족하는 Sparse한 비결정론적 MDP의 경우에도 비슷한 방법으로 풀이 시간이 $\max_i GD(s,s_{r_i})$ 에 비례하는 bound를 가짐을 보일 수 있다.

즉 다중 보상 MDP의 경우 단일 보상 MDP와 다르게 평균 측지 거리가 아닌 최대 측지 거리에 의해 수행 시간이 bounded 되게 된다.

4. 결론

본 논문에서는 MDP의 상태 변화 그래프의 위상적 특성과 그 풀이 효율에 대한 기존의 실험적 연구를 바탕으로, MDP의 상태 변화 그래프의 위상적 성질인 평균 측지 거리와 그 MDP를 평가치 반복과 Q-learning로 풀 경우의 풀이 시간간의 관계에 대해 분석해 보았다.

그 결과 몇 가지 가정 하에, 단일 보상 MDP의 경우 상태 변화 그래프의 평균 측지 거리에 비례하는 bound를 갖고, 다중 보상 MDP의 경우 상태 변화 그래프의 최대 측지 거리에 비례하는 bound를 갖는다는 것을 보일 수 있었다. 하지만 일반적으로 낮은 평균 측지 거리를 갖는 작은 세상 그래프 모델의 경우 낮은 최대 측지 거리도 갖게 되고, 이는 평균 측지 거리와 풀이 성능 간에 실험적으로 나타나는 높은 상관관계를 설명해 준다.

향후 과제로는 좀더 Tight한 bound의 확립과 일반적인 다중 보상 MDP의 경우에 대해서 측지 거리의 분포를 고려한 추가적인 분석, 그리고 이러한 분석을 활용한 실세계에 적용 가능한 효율적인 MDP 디자인이 있을 것이다.

감사의 글

이 논문은 과학기술부 국가지정연구실사업(NRL)에 의하여 지원되었음.

참고 문헌

[1]Barto, A.G., Mahadevan, S. Recent advances in hierarchical reinforcement learning. Discrete Event Systems Journal, 13, 41-77, 2003.

[2]Dietterich, T. G. (1998). Hierarchical reinforcement learning with the MAXQ value function decomposition. In Proceedings of the 15th International Conference on Machine Learning ICML'98.

[3]Sutton, R.S., Precup, D., Singh, S.P. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence, 112, 181-211, 1999.

[4]John Rust, 1997. A Comparison of Policy Iteration Methods for Solving Continuous-State, Infinite-Horizon Markovian Decision Problems Using Random, Quasi-random, and Deterministic Dircritizations, Computational Economics 9704001, EconWPA.

[5]Amy McGovern , Andrew G. Barto, Automatic Discovery of Subgoals in Reinforcement Learning using Diverse Density, Proceedings of the Eighteenth International Conference on Machine Learning, p.361-368, June 28-July 01, 2001

[6]J. Pineau, G. Gordon, and S. Thrun, Policy-contingent abstraction for robust robot control, Conference on Uncertainty in Articifical Intelligence (UAI), August, 2003, pp. 477 - 484

[7]B. Digney, "Learning Hierarchical Control Structure for Multiple Tasks and Changing Environments," Proceedings of the Fifth Conference on the Simulation of Adaptive Behavior: SAB 98, 1998

[8]Mannor, S., Menache, I., Hoze, A., & Klein, U. (2004) Dynamic abstraction in reinforcement learning via clustering. ICML, 21: 560--567. 13

[9]Sutton, R.S. Barto, A.G. Reinforcement learning: an introduction. MIT press, 1998.

[10]Watts, D.J., Strogatz, S.H. Collective dynamics of 'small-world' networks. Nature, 393, 404-407, 1998

[11]Barabasi, A.-L., Albert, R. Emergence of scaling in random networks. Science, 286, pp. 509-512,. 1999.

[12]L. da F. Costa a; F. A. Rodrigues a; G. Travieso a; P. R. Villas Boas a. Characterization of complex networks: A survey of measurements, Advances in Physics, Volume 56, Issue 1 January 2007 , pages 167 - 242

[13]이승준, 장병탁. 복잡계의 위상특성을 이용한 MDP 학습의 효율 분석, 한국정보과학회 가을학술발표 논문집,