論 文
52D-9-4

임무수행을 위한 개선된 강화학습 방법

An Improved Reinforcement Learning Technique for Mission Completion

權 又 英* · 李 相 勳** · 徐 一 弘***
(Wooyoung Kwon · Sanghoon Lee · Il Hong Suh)

Abstract - Reinforcement learning (RL) has been widely used as a learning mechanism of an artificial life system. However, RL usually suffers from slow convergence to the optimum state-action sequence or a sequence of stimulus-response (SR) behaviors, and may not correctly work in non-Markov processes. In this paper, first, to cope with slow-convergence problem, if some state-action pairs are considered as disturbance for optimum sequence, then they are to be eliminated in long-term memory (LTM), where such disturbances are found by a shortest path-finding algorithm. This process is shown to let the system get an enhanced learning speed. Second, to partly solve a non-Markov problem, if a stimulus is frequently met in a searching-process, then the stimulus will be classified as a sequential percept for a non-Markov hidden state. And thus, a correct behavior for a non-Markov hidden state can be learned as in a Markov environment. To show the validity of our proposed learning technologies, several simulation results will be illustrated.

Key Words : reinforcement learning, delayed reward, markov process, batch process

1. 서 론

지능형 에이전트(intelligent agent)가 동적으로 변화하는 환경에서 주어진 목적을 달성하기 위해서는, 상황에 가장 적절한 행동을 시행착오(trial-and-error)를 통해 환경과 상호작용하면서 학습해야 한다. 이를 위해 동물의 학습을 응용한 학습방법이 널리 이용되고 있다. 민스키(Minsky)에 의해 소개된 강화학습(Reinforcement learning)은 상태에 대한 행동의 규칙을 보상신호를 통하여 학습해 나가는 방법으로 환경에 대한 정확한 사전지식이 없이 학습 및 적응성을 보장할 수 있어서 지능형 에이전트의 학습방법으로 유용하다[1]. 그러나 기존 강화학습을 실제 문제에 적용할 경우 몇 가지 문제점이 나타난다. 첫 번째로, 실제 환경에서는 목표에 도달할 때까지는 중간 단계의 행동에 대해서 즉각적인 보상이 주어지지 않는다. 이러한 지연보상문제에 대한 일반적인 해결방안은 보상신호를 과거의 상태와 행동에 배분하는 일시적인 신뢰할당을 하는 것이다. 그러나 이러한 방법은 최적의 상태와 행동을 찾아내는 수렴속도가 느리다는 문제를 갖는다. 두 번째로 일반적인 강화학습 방법은 비 마르코프 환경(non-Markov Environment)에서 적절히 동작하지 않는 문제가 있다. 비 마르코프 환경은 현재 상태가 과거의 상태와 행동에 의해 영향을 받는 환경이다. 이러한 비 마르코프 환경은 환경의 변화가 복잡하거나, 현재 상태를 인지하는 감각기의 능력이 부족할수록 나타나기 쉽다.

본 논문에서는 지연 보상 문제에서 기존 강화학습 방법에 비해 빠른 수렴속도를 보이며, 비 마르코프 환경에서 동작할 수 있는 임무수행 학습 방법을 제안한다. 2장에서는 임무수행 학습과 비 마르코프 환경의 개념을 서술하고, 그에 대한 기존 연구방법들을 소개하였다. 3장에서는 지연 보상 문제에 대해 빠른 수렴속도를 보이며, 비 마르코프 환경에서 동작하는 새로운 학습 방법을 제안한다. 본 학습방법은 지연 보상 문제에서의 강화학습 방법으로 탐색 된 상태공간에서 내부 추론과정을 적용하여 목표에 도달하는데 불필요한 상태-행동 규칙을 삭제한다. 모든 상태에서 목표 상태까지 도달하기 위한 최단경로를 찾는 방법으로 다익스트라(Dijkstra)의 알고리즘을 사용한다. 또한 제안하는 학습방법에서는 현재의 상태를 과거의 상태-행동 규칙들과 조합하여 시계열 자극을 구성하며, 이를 통해 비 마르코프 은닉 상태를 구분할 수 있도록 한다. 4장에서는 제안한 학습 방법의 타당성을 검증하기 위하여, 2차원 격자환경에서의 실험을 하였다. 시작지점에서 목표지점가지 찾아가는 행동의 순서를 학습하는 과정을 마르코프 환경과 비 마르코프 환경에서 수행하였으며, 이를 Q-학습 방법과 비교하여 수렴속도가 향상됨을 보이고, 비 마르코프 환경에서도 적절히 동작함을 보였다.

* 正 會 員 : 漢陽大學 情報通信大學院 碩士
** 正 會 員 : 漢陽大學 電子電氣制御計測工學科 博士課程
*** 正 會 員 : 漢陽大學 情報通信大學院 正敎授
接受日字 : 2003年 3月 26日
最終完了 : 2003年 7月 18日

2. 연구배경

2.1 행동순서의 학습

마르코프 환경(Markov Environment)은 이전 상태와 이에

대한 행동들이 현재 상태를 지배하는 환경이며, 이에 대한 수학적 표현은 (1)과 같다.

$$\Pr(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, s_0, a_0) = \Pr(s_{t+1} \mid s_t, a_t) \quad (1)$$

여기서 st는 시간 t에서의 상태이며, at 는 시간 t에서 선택된 행동이다. 또한 t-1과 t+1은 각각 과거와 미래를 표현한다. 환경이 마르코프적인 경우, 즉 식(1)을 만족하면 과거의 자극과 행동에 대해 파악할 필요가 없으며, 바로 이전 상태와 그에 대한 행동만을 관측한다면 현재상태를 파악 할 수 있다. 따라서 연속적인 행동들은 자극에 대한 반응의 연쇄로 설명 될 수 있다. (그림 2-5)에서, 만약 에이전트가 s1-a1, s2-a2, s3-a3 에 대한 자극과 반응의 관계를 알고 있다면, 각 자극에 대해 적합한 행동을 하는 것으로 s1에서 s4까지 도달하는 행동의 순서를 만들어 낼 수 있다.

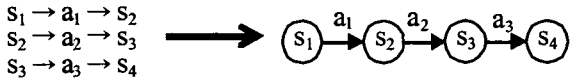

그림 1 자극과 반응의 연쇄
Fig. 1 An example representing a behavior sequence

결국 행동의 순서(behavior sequence) 는 자극에 대한 행동들의 연쇄로 표현되며 이는 결국 유한상태기계(FSM : Finite State Machine)로 나타낼 수 있다.

그러나 대부분의 실제 환경은 위에서 언급한 마르코프적인 요소를 모두 만족하지 않는다. 현재의 상태를 이전 상태들과 이에 대한 행동만으로 표현 할 수 없는, 즉 자극과 반응의 연쇄로 현재의 상태를 표현할 수 없는 상태를 비 마르코프 은닉 상태(non-Markov hidden state)라고 한다. 이에 대한 수식은 (2)에 나타나 있다.

$$\Pr(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, s_0, a_0) \neq \Pr(s_{t+1} \mid s_t, a_t) \quad (2)$$

비 마르코프 환경에서는 동일한 외부자극을 인식할 경우라 하더라도 과거의 행동들에 따라서 현재의 상태가 달라 질 수 있다. 결국 비 마르코프 은닉 상태는 현재 상태가 외부 자극 뿐 아니라 다른 요소-특히 과거의 행동들-에 의해 영향 받지만, 외부 자극만으로는 구분해 낼 수 없는 은닉 상태로 볼 수 있다. 따라서 비 마르코프 은닉 상태를 구분해 내 즉 현재 상태에 대한 판단을 외부자극에만 의존하는 것이 아니라 외부 자극과, 과거의 상태와 행동을 기록한 내부자극의 조합에 의존하여 판단한다면, 비 마르코프 은닉 상태는 인식할 수 있는 상태가 된다.

이를 위한 가장 간단한 방법은 과거의 자극과 행동 모두를 조합하는 것이지만 이 방법은 계산 복잡도를 크게 증가시킨다. 따라서 관건은 과거의 자극과 행동들의 조합 중 중요한 것만을 뽑아내는데 있다. 기존 연구들로서 순서에 어떻게 조합하느냐로 볼 수 있으며, 은 비 마르코프 은닉 상태를 인식하여 학습하는 방법은 은닉 상태를 어떻게 인식하느냐로에는 여러 가지 방법이 소개되어 있다[2][3][4][5].

2.2 S-R학습과 강화학습

강화학습은 자극과 행동의 관계를 학습하는 방법으로서 가장 많이 사용되고 있다. 강화학습 방법은 주어진 상태에서 보상을 최대화하는 행동의 규칙을 찾아내는 방법이며, 학습자는 각각의 행동에 대해 즉각적인 보상을 주지 않고 최종적인 결과에 대해서만 보상을 준다. 따라서 최종 결과에 대해서만 평가를 하는 지연 보상 문제에서도 적용 가능한 방법이다. 지연 보상의 경우에 목적을 달성하기 위한 자극과 행동에 대해 어떻게 신뢰할당을 할 지 여부는 강화학습의 중요한 문제이다[6][7].

강화학습과 도구적 조건화(operant conditioning)[8]는 둘 다 자극과 행동간의 관계를 학습한다는 점과, 학습자가 목적을 달성하기 위한 모든 자극과 행동에 대해 보상을 주지 않고 그 결과에 대해서만 보상을 준다는 점에서 유사성을 가진다[9]. TD 학습방법과 Q-학습방법은 강화학습의 가장 일반적인 구현 방법이다. 강화학습은 위에서 언급한 마르코프 환경에서 행동의 순서를 학습하는데 적합한 방법이다. 그러나 일반적인 강화학습은 지연 보상의 경우에 느린 수렴속도를 보이며, 상태공간이 커지면 이런 현상은 더욱 심각해진다[10]. 또한 강화학습은 비 마르코프 환경에서는 적절히 학습하지 못한다. 이에 대한 고전적인 해결책은 환경에 대한 정확한 정보를 에이전트에게 전달해주는 방법이다[11]. 그러나 외부 환경에 대한 모델을 만드는 것이 불가능한 경우에는 이러한 모델 기반 접근 방법은 유효하기 않다. 따라서 모델에 독립적인 접근방법이 요구되며, 그 방법 중 하나는 계층적인 강화학습 방법이다. 계층적 강화학습 방법은 비 마르코프적인 임무를 마르코프적인 세부 임무들로 분할하여 각각을 기존 강화학습 방법으로 해결하는 접근 방법을 사용한다[12][13][14]. 또 다른 모델 독립적인 방법은 비 마르코프 은닉 상태를 구분해 내는 접근 방법이다. 비 마르코프 은닉 상태는 과거의 자극과 행동에 관련되어 있기 때문에, 과거의 자극과 행동을 저장하는 단기기억을 사용하여 비 마르코프 은닉 상태를 구분할 수 있다[2][15][16].

3. 최단경로 탐색 방법을 포함하는 빠른 강화학습

인공생명체는 최대한 적은 시도를 통해 자극과 행동의 관계를 학습해야 한다. 일반적으로 학습에 필요한 탐색과정은 많은 비용이 요구되는 과정이며, 때때로 시스템을 위험한 상황에 처하도록 한다. 반면에 내부적인 추론과정은 계산복잡도를 증가시키지만 탐색과정에 비해 비용을 적게 할 수 있다는 장점이 있다[2].

인공생명체의 학습방법으로 널리 사용되는 강화학습은 보상을 받는 시점에서부터 점진적으로 신뢰할당이 이루어진다. 따라서 학습해야 하는 하나의 행동순서의 신뢰도는 서서히 높아지고, 이 때문에 비교적 느린 수렴속도를 보인다. 게다가 행동순서의 학습을 위한 자극과 행동이 많아지면 상태공간이 증가되어 수렴속도는 더욱 느려진다.

본 장에서는 내부적인 추론 과정을 사용하여 불필요한 자극과 행동의 신뢰도를 감소시키는 방법을 적용하여 빠른 수렴속도를 보이는 새로운 강화학습 방법을 제안한다. 먼저 강화학습으로 외부 환경에서 주어진 자극에 대한 행동들을 학습한다. 하나의 에피소드를 구성한 자극과 행동들은 모두 단

기기억(STM : Short Term Memory) 공간에 저장되며, 이들은 신뢰할당 과정을 거쳐서 장기기억(LTM : Long Term Memory)으로 전환된다. 여기까지는 일반적인 강화학습 방법과 동일하지만, 그 후 장기기억에 저장된 자극과 행동들에 대해 내부 추론과정을 사용하여 수렴속도를 빠르게 한다. 장기기억에 저장된 내용들을 대상으로 모든 상태에서 목적상태에 도달하는 최적경로를 구하고, 이에 해당하지 않는 요소들을 장기기억에서 제거한다. 위의 과정을 거친 장기기억에는 목적상태에 도달하기 위한 최단 경로들만이 저장되게 된다. 내부 추론과정에 의해 제거된 장기기억 요소는 결국 목적에 도달하는데 불필요한 요소들이다.

3.1 강화학습 방법을 이용한 탐색과정

제안된 학습방법에서 주어진 상태와 실행한 행동을 저장하는 내부 기억공간은 단기기억과 장기기억으로 구성된다. 단기기억은 각 시간에서 선택된 자극과 그에 대한 행동을 시간 순서에 따라 기억하며, 장기기억은 자극과 행동에 대한 신뢰도를 저장한다. 단기기억에서 장기기억으로 전환되는 과정에서 해당 상태에 대한 행동의 중요도를 결정하는 신뢰 할당이 이루어지며, 그 과정은 (그림 2)에 나타나 있다.

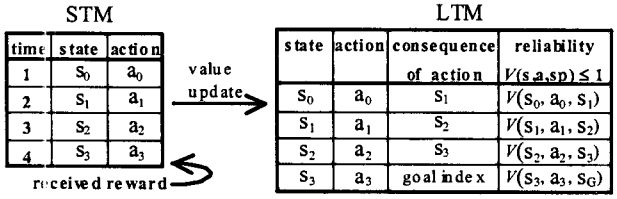

그림 2 단기기억에서 장기기억으로 전환되는 과정
Fig. 2 STM and LTM

신뢰할당이 중복되는 것을 방지하기 위하여, 신뢰할당이 이루어진 후의 단기기억은 초기화되어 새로운 자극과 행동을 기록한다. 어떤 행동의 결과로 보상을 받게 된다면 이전에 했던 행동들, 즉 단기기억에 저장된 모든 내용은 보상을 받기 위한 사전행동으로 볼 수 있다. 단기기억에 저장된 상태-행동규칙중에서, 최근의 상태-행동규칙은 이전 상태에 비해 중요도가 크다고 가정할 수 있다. 따라서 신뢰할당은 현재상태에 가까운 상태-행동규칙들에 높은 가중치를 부여하며, 이러한 사항은 (3)에 나타난 신뢰도 값 V에 반영된다.

$$V(s_i,a_i,s_{i+1}) \leftarrow V(s_i,a_i,s_{i+1}) + \frac{\eta(1-v(s_i,a_i,s_{i+1}))}{(N-i+1)^{\lambda}} \quad (3)$$

여기서 si와 ai는 각각 시간 t에서의 자극과 행동이다. 그리고 η는 학습비율 λ는 감쇄율, N은 단기기억에 저장된 자극과 행동의 개수 그리고 (N-i)λ는 보상을 받은 시간에서부터의 거리에 따른 가중치이다. (3)에 표현된 수식은 강화학습 방법 중 몬테-카를로 방법(Monte-Carlo method)과 유사하다. (그림 3)은 강화학습 방법에 의해 단기기억과 장기기억이 갱신되는 과정을 나타낸다.

본 논문에서 사용하는 신뢰도 V(s,a,sp)는 Q-학습방법의 Q(s,a)와 동일한 역할을 하며, 갱신된 신뢰도 V는 새로운 탐색과정에서 자극과 행동을 선택하는데 확률적으로 영향을 준다. 신뢰도를 기준으로 자극과 행동을 선택하기 위하여 수정된 볼츠만 탐색방법(Boltzman exploration)이 사용되었으며 그 식은

$$\Pr(s,a) = \frac{e^{\frac{V(s,a,sp)}{T}}}{\sum_{a'}\sum_{s'} e^{\frac{V(s',a',sp)}{T}}} \quad (4)$$

와 같다.

본 논문에서 사용된 신뢰도인 V값은 Q-학습방법에서 사용된 Q값과는 달리 3개의 인자를 가진다. 세 번째 인자인 sp는 현재 선택된 자극과 행동의 예측 결과를 나타낸다. 이 세 번째 인자를 사용함으로써 추론과정을 손쉽게 할 수 있으며, 또한 다음 장에서 설명할 비 마르코프 은닉 상태를 구분하는데 효과적으로 이용될 수 있다.

repeat(for each episode)
  initialize STM
  repeat (for each step of episode)
    choose action and state a,s with respect to $V(s,a,sp)$ (Boltzman exploration)
    store a, s to STM
    take action a
  LOOP i = 1 to STMsize
    IF i=STMsize
      $V(s_i,a_i,\text{Goal Index})=1$
    ELSE
      $V(s_i,a_i,s_{i+1}) \leftarrow V(s_i,a_i,s_{i+1}) + \eta \frac{1-V(s_i,a_{i-1},s_{i-1})}{i^{\lambda}}$
$V(s,a,sp)$: LTM Value
sp : consequence of action
η : learning rate $0<\eta<1$
λ : decay rate $0<\eta<1$

그림 3 단기기억과 장기기억 갱신과정의 의사코드
Fig. 3 Psedo code of suggested learning algorithm

3.2 추론에 의한 최단경로 탐색

내부 추론과정은 내부 기억공간을 최적화 시켜, 탐색과정에서 최적의 자극과 행동을 선택하는데 도움을 주는 과정이다. 장기기억 공간에는 모든 상태에서 목적 상태에 도달하기 위한 경로를 포함하고 있다. 자극-행동-행동의 결과로 이루어진 장기기억 요소들을 연결하면 유한상태기계(FSM : Finite State Machine)를 구성할 수 있다. 이렇게 구성된 유한상태 기계에는 탐색한 모든 상태에 대한 정보가 포함되어 있으며, 이중에는 목적상태에 도달하기 위해 불필요한 요소들이 포함되어 있다. 이러한 불필요한 요소들을 내부 추론과정에 의해 삭제하고, 최적 경로의 신뢰도를 강화해준다면 학습의 수렴속도를 향상시킬 수 있다.

(그림 4)에는 장기기억 요소들을 유한상태기계로 표현한 예를 보이고 있다. 유한상태기계는 방향성 그래프로 표현될 수 있다. 따라서 최단 경로를 탐색하기 위해 그래프 이론에서 소개한 최단경로 탐색 기법을 사용한다.

(그림 4a)는 전체 상태공간에 대한 유한상태기계(또는 방향성 그래프)를 나타내며, (그림 4b)는 그 중에서 최단경로만을 표현한 것이다. 최단경로 탐색 기법으로는 다익스트라

의 알고리즘이 사용되었다[17]. 최단경로 탐색이 수행된 후, 최단경로에 포함되지 않은 장기기억 요소들은 불필요한 것으로 간주되어 삭제된다.

(그림 5)는 최단경로 탐색 방법을 포함하는 내부 추론과정의 의사 코드이다.

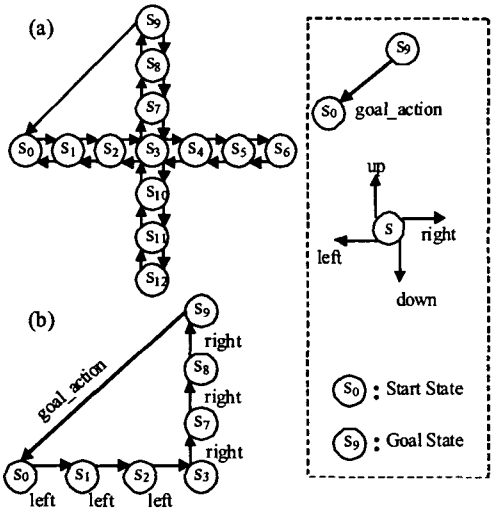

그림 4 FSM에서 최단경로 탐색의 예

Fig. 4 An example of FSM to find the shortest path

```
Create Graph G(V,E)
repeat for all l ∈ LTM
  get s,a,sp from LTM entry l
  G(V,E) ← l(s start Vertex, a as Edge sp as end Vertex, weight=1)
end

find all pair shortest path P from G using Djkstra's algorithm

repeat for all l ∈ LTM
  get s,a,sp from LTM entry l
  make V,E from l
  if V,E ∈ G then
     V(s,a,sp) ← V(s,a,sp)+α
     clamp(V(s,a,sp),0,1)
  else
     V(s,a,sp) ← γ×V(s,a,sp)
end
```

그림 5 내부 추론 과정의 의사 코드

Fig. 5 Pseudo code the resoning process

3.3 제안된 방법의 최적성과 수렴성

일반적인 몬테 카를로 방법은 그 최적성과 수렴성이 초기 탐색과정에 크게 의존하기 때문에 엄밀히 검증되어 있지 않다. 제안된 학습 방법은 몬테 카를로 방법에 기초하고 있기 때문에 동일한 문제가 발생하게 된다. 그러나 제안된 학습방법에서는 내부 추론 과정에 의해 수렴속도가 향상될 뿐 아니라, 수렴성과 최적성도 크게 나아지게 된다.

제안된 학습방법은 탐색과정의 모든 상태공간에 대한 정보를 장기기억에 저장한다. 또한 내부 추론과정은 이렇게 저장된 상태공간을 대상으로 모든 상태에서 목적상태에 도달하는 최적 경로를 탐색한다. 따라서 장기 기억이 최적 경로를 포함하게 된다는 것을 나타낼 수 있다면 제안된 학습방법의 최적성과 수렴성을 밝힐 수 있다.

이를 위해 두가 경우를 가정해 볼 수 있다. 첫째로, 장기기억이 최적경로를 포함하고 있는 경우이다. 이 경우에는 최적경로에 대한 신뢰도의 증가가 내부 추론과정에 의해 계속적으로 이루어지기 때문에 결과적으로 최적 경로로의 수렴이 이루어진다. 두 번째로 장기기억이 최적경로를 포함하고 있지 않은 경우를 가정해 볼 수 있다. 이 경우 장기기억에 저장된 상태공간내의 부분 최적 경로만이 발견될 수 있다. 그러나 내부추론 과정에 의해 부분 최적 경로를 구성하는 상태-행동규칙의 신뢰도가 증가한다 하더라도, 아직 탐색해보지 않은 상태 공간을 탐색할 가능성은 적지만 항상 존재한다. 따라서 탐색횟수가 무한히 증가한다면, 장기기억은 최적 경로를 포함하게 된다. 이 경우 최적경로를 구성하는 상태-행동규칙에 대한 신뢰도가 내부 추론과정에 의해 증가하게 되므로, 최종적으로 에이전트는 최적경로로 수렴하게 된다. 탐색과정의 가중치τ,학습비율 η 와 내부 추론과정의 변수 α를 조절함으로써, 수렴속도와 최적성의 관계를 조절 할 수 있다.

4. 비 마르코프 환경에서의 학습방법

비 마르코프 은닉 상태를 구분해 내기 위해서는 과거의 자극에 대한 행동의 기록을 참조하여 현재의 상태를 재구성 해야한다. 이를 위해 과거의 자극과 행동이 시간 순서에 따라 저장된 단기기억을 활용한다. (그림 6)은 비 마르코프 은닉 상태를 포함하는 미로 환경을 나타낸다. (그림 6a)에서 각 셀의 번호는 에이전트 주변의 센서 정보에 따른 상태를 나타낸다. 이 환경에서 에이전트는 시작상태 1에서 목적상태 G까지 도달해야 한다. (그림 6c)는 시작상태에서 목표상태까지 도달하는 탐색과정의 예를 나타내고 있다. 첫 시도에서는 신뢰도가 모두 같으므로 임의의 경로를 거쳐 목적상태에 도달한다. 목적상태에 도달하게 되면 자극 s2에 대해 행동 up과 down이 동시에 신뢰할당을 받지만, s2-down이 목적상태와 가깝기 때문에 더 높은 신뢰도를 갖는다. 따라서 두 번째 시도에서는 s1과 s2사이를 반복하는 행동을 되풀이한다. (그림 6b)는 s2에 대한 유한 상태기계 정보를 나타낸다. 여기에서 자극 s2에 대해 두개의 행동(up 과 down)이 모두 적합한 행동이 된다. 시작상태에 가까운 자극 s2에 대해서는 행동 up이 적합하며 목표상태에 가까운 자극 s2에 대해서는 행동 down 이 적합하다. 그러나 이 경우에는 서로 다른 자극 s2를 구분하지 못하기 때문에, 자극 s2에 대한 올바른 행동을 선택할 수 없다. 이러한 현상은 비 마르코프 은닉 상태인 s2를 구분해 내지 못하는 데서 발생하는 문제이며, 본 연구에서는 현재 자극과 과거의 자극-행동들을 조합한 새로운 자극을 생성하여 이 문제를 해결한다.

(그림 7)에서처럼 자극 s1에 대해 행동 up을 했을 경우 나타나는 현재의 자극 s2를 s2a라고 정의하고, 자극 s3에 대해 행동 down을 했을 경우 나타나는 자극 s2를 s2b라고 정의한다면, 비 마르코프 은닉 상태를 구분해 낼 수 있다. 이것은 '자극-행동→행동의 결과' 로 표현되며, 이를 시계열 자극이라고 정의한다. 두개 이상의 연속된 시계열 자극은 (그림 7)에서처럼 시계열 자극간의 연쇄로 설명될 수 있다.

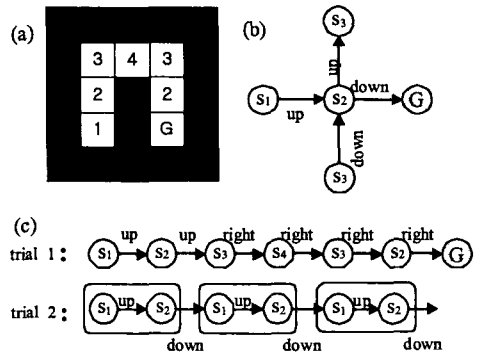

그림 6 비 마르코프 은닉 상태를 갖는 환경의 예
Fig. 6 An example of non-Markov hidden state

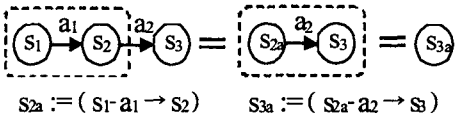

그림 7 시계열 자극의 연쇄과정
Fig. 7 Chaining Sequential Perception

모든 자극을 시계열 자극으로 판단하는 것은 비효율적이므로, 비 마르코프 은닉상태와 관련 있는 자극만을 시계열 자극으로 만드는 것이 바람직하다. 본 연구에서는 탐색과정 중 정해진 횟수 이상 반복하여 나타나는 자극과 행동을의 집합에 대해서 새로운 시계열 자극을 생성한다.

시계열 자극을 생성하는 것은, 에이전트가 인식할 수 있는 상태를 증가시키는 것이므로 학습 속도를 느리게 한다. 그러나 내부적인 추론과정을 통해 최적의 상태-행동 규칙을 찾아내면 시계열 자극의 생성으로 상태가 증가하더라도 빠른 수렴속도를 보일 수 있다. (그림 8)은 시계열 자극을 생성하는 과정에 대한 의사 코드이다.

위의 모든 과정을 포함하는 전체 학습과정의 의사 코드는 (그림 9)에 나타나 있다.

```
N=SizeSTM
loop i=1 to N-1
{
  get s1,a1 from index i
  get s2,a2 from index i+1
  count=0
  loop j=i to N-1
  {
    get s1',a1' from index j
    get s2',a2' from index j+1
    if ( s1 == s1' and a1 == a1' and s2 == s2')
      count=count+1
  }
  if ( count > threshold )
    createNewSequencialPercept(s1,a1,s2)
}
```

그림 8 시계열 자극을 생성하는 과정의 의사 코드
Fig. 8 Pseudo code to generate Sequential Perception

```
repeat for each episode
{
  initialize ( STM )
  repeat for each step of episode
  {
    generateSeqentialPercepts()
    selectAction()
  }
  updateLTM()
  reasoningProcess()
}
```

down down

그림 9 전체 학습과정
Fig. 9 Overall learning algorithm

(그림 8)에 나타난 방법은 즉 시계열 자극이 인식되면 특별한 신뢰할당 과정 없이 빈도수에 따라서 즉각적으로 새로운 상태를 생성하기 때문에, 기존의 장기기억 공간에 새로운 상태를 추가하는 역할만을 한다. 즉 새로운 상태를 학습하는 데에 Instant-based 학습 방법을 사용하고 있다. 이에 비해 기존의 계층적 강화학습[2][3][5]은 시계열 자극을 시계열 자극을 계층적으로 구성하며 이에 대해 각각 신뢰할당을 한다. 시계열 자극을 계층적으로 구성하는 방법은 비교적 길이가 긴 시계열 자극을 효과적으로 처리할 수 있으며 저장 공간을 적게 차지한다. 따라서 필요한 시계열 자극만을 골라내는데 더 효과적이다. 그러나 시계열 자극에 대한 신뢰할당이 점진적으로 이루어지기 때문에 수렴속도가 매우 늦다. 본 연구에서는 학습 방법의 수렴속도에 중점을 두고 있기 때문에 인식된 시계열 자극을 즉각적으로 새로운 상태로 인식하도록 하는 방법을 사용하였다.

5. 실 험

제안한 학습 알고리즘의 타당성을 검증하기 위하여, 두개의 실험을 수행하였다. 첫 번째 실험에서는 마르코프 환경을 가정한 (그림 10)과 같은 H형태의 미로를 대상으로 하였으며, 두 번째 실험에서는 비 마르코프 환경을 가정한 (그림 12)와 같은 M형태의 미로 환경을 대상으로 하였다. 각 미로에서 에이전트는 시작상태 S에서 목적상태 G까지 도달하기 위한 최적의 경로를 학습하여야 한다. 또한 에이전트는 각 상태에서 다음 상태로 가기 위하여, 위, 아래, 오른쪽, 왼쪽 방향으로의 행동을 선택 할 수 있다.

5.1 마르코프 환경에서의 학습

본 실험에서는 (그림 10)과 같은 H형태의 환경에서 실험하며, 에이전트에 입력되는 상태정보는 각각의 위치정보 이다. 따라서 상태정보 들은 중복되지 않으며, 실험 환경은 마르코프 환경이다. (그림 11)은 Q-학습방법과 내부추론과정을 포함한 제안된 알고리즘과의 비교 결과를 표현한다. (그림 11)에서 Q-학습방법은 28회의 시도만에 최적 경로인 15개의 상태전이 횟수로 수렴하는 것을 볼 수 있다. 그러나 제안한 알고리즘은 4회의 시도만에 유사 최적 경로(sub-optimal)인 20개의 상태전이 횟수로 수렴함을 알 수 있다. 따라서 제안한 학습 방법이 수렴속도를 크게 향상시킴을 알 수 있다.

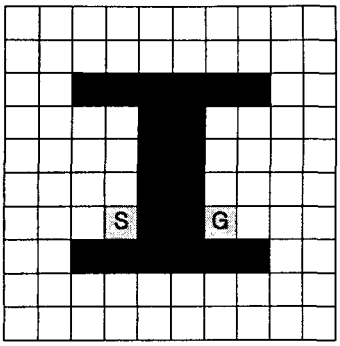

그림 10 H 형태의 격자환경
Fig. 10 H type maze

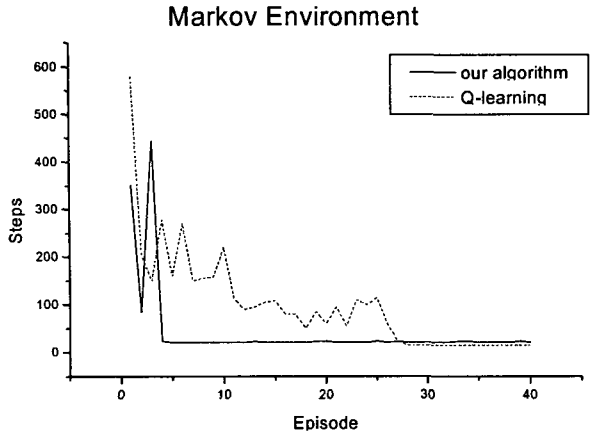

그림 11 H 형태의 격자환경에서의 실험결과
Fig. 11 Experiment results for H type maze

5.2 비 마르코프 환경에서의 학습

(그림 12)에서 나타난 M형태의 미로에서 비 마르코프 환경에서의 학습능력을 검증한다. 에이전트로 입력되는 외부 환경정보는 주변의 센서정보만으로 구성되며, (그림 12)에서 보이는 각 셀의 번호는 주변 센서정보에 따른 특징을 표현한다. 예를 들어 2번 셀에 위치한 에이전트는 왼쪽과 오른쪽으로 가는 행동이 차단되어 있다. 이 정보만으로는 현재의 위치를 정확히 판정하는데 어려움이 있으며, 따라서 상태 2는 비 마르코프 은닉 상태로 볼 수 있다. 제안한 학습 방법은 4자에서 설명한 바와 같이 이러한 비 마르코프 은닉 상태를 시계열 자극으로 구성하여 구분해낸다. 시계열 자극의 길이가 길어질수록 수렴 속도 및 안정성은 크게 떨어진다. (그림 12)의 환경에서는 시계열 자극의 길이를 3회로 제한할 경우에 가장 좋은 성능을 보였다.

(그림 12)와 같은 환경에서 Q-학습방법을 실험하여 보았으나 수렴하지 않음을 알 수 있었다. 첫 번째 시도에서는 많은 상태전이 끝에 목표에 도달하였다. 첫 번째 골에서 가장 가까운 자극이 상태 2이므로 자극 2에 대해서 아래로 가는 행동이 가장 큰 신뢰도를 갖게 된다. 다음 시도에서 에이전트는 상태 S에서 위로 가는 행동과 상태 2에서 아래로 가는 행동을 반복하게 된다. 결국 에이전트는 목표상태에 도달하지 못하고 같은 구간을 계속 반복하여 이동한다. 이 예는 Q-학습방법이 비 마르코프 은닉 상태를 효과적으로 다루지 못함을 보이고 있다.

제안한 알고리즘의 실험 결과는 (그림 13)에 나와 있다. (그림 13)은 10회의 실험에 대한 평균적인 상태전이 수를 표현하고 있으며, 각각의 실험은 모두 100회의 시도로 구성되어 있다. 제안한 학습 방법이 세 번째 시도 이후에 급격히 현저히 낮은 상태전이 횟수를 보이는데, 이는 내부적인 추론과정에 의한 영향으로 볼 수 있다. 그러나 항상 최적의 상태전이 횟수로 수렴하지는 않음을 알 수 있다. 또한 몇몇 시도에서는 갑자기 많은 상태전이 횟수를 보이는 것을 볼 수 있다. 이에 대한 원인은 시계열 자극이 상태의 수를 크게 증가시키기 때문이다. 다른 원인은 불필요한 시계열 자극이 생성되기 때문이다. 예를 들어 (6-down)-(2-down)-2는 목적상태에 도달하는데 꼭 필요한 시계열 자극이다. 그러나 (2-down)-(2-left)-2는 미로 상에서 빈번히 나타날 수 있으며, 이는 목표에 도달하는데 불필요한 시계열 자극이다.

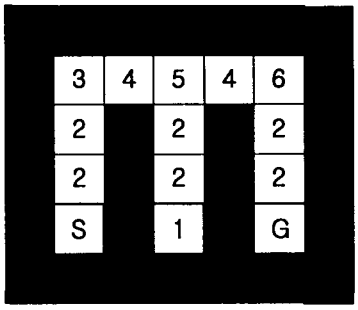

그림 12 M 형태의 미로
Fig. 12 M type maze

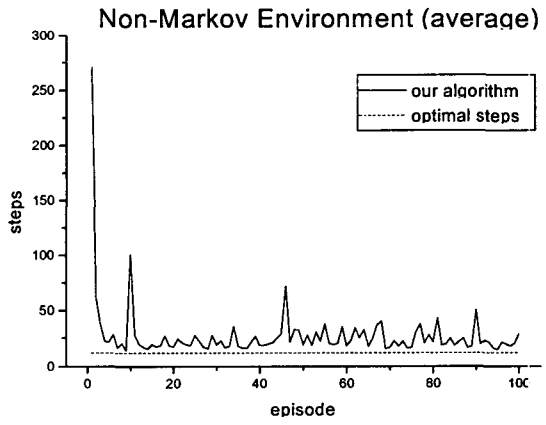

그림 13 M 형태의 미로에서 학습결과
Fig. 13 Learning at non-Markov Environment(average)

6. 결 론

본 논문에서는 행동의 순서를 효율적으로 학습하기 위한 개선된 강화학습 방법을 제안하였다. 첫째로 강화학습 방법으로 탐색 된 상태공간을 대상으로 최적경로 탐색기법을 사

용하여 최적경로가 아닌 상태-행동규칙을 장기기억공간에서 삭제하는 방법을 적용하였다. 이 과정은 제안한 학습 방법의 수렴속도를 크게 개선하였다. 둘째로 단기 기억공간을 활용하여 현재 자극과 과거의 상태-행동을 조합한 시계열 자극을 구성하는 방법으로 비 마르코프 은닉 상태를 갖는 자극을 구분해 낼 수 있도록 하였다. 본 논문에서 제안한 학습방법을 검증하기 위하여, 마르코프 환경을 갖는 H형태의 격자 환경과, 비 마르코프 환경을 갖는 M형태의 미로에서 각각 실험을 하였으며 각각 수렴속도의 향상과, 비 마르코프 환경에서 동작함을 확인할 수 있었다.

참 고 문 헌

[1] M.L. Minsky, "Steps towards artificial intelligence", In Proceedings of the Institute of Radio Engineers, 49, pp8-30, 1961.

[2] A. K. McCallum, "Reinforcement Learning with selective Perception and Hidden State", PhD thesis, University of Rochester, 1996.

[3] R. Sun, C. Sessions, "Self Segmentation of Sequences", IEEE Trans System Man and Cybernetics, vol. 30, no. 3, pp.403-418, 2000.

[4] M. L. Littman, "Algorithm for Sequential Decision Making", PhD thesis, Brown University, 1996.

[5] S. D. Whitehead, L.J. Lin, "Reinforcement learning in non-Markov environments", Artificial Intelligence, 1993.

[6] R. Sutton, A. Barto, Reinforcement Learning, MIT Press, 1997.

[7] C. Watkins, "Learning from Delayed Rewards", PhD thesis, University of Cambridge, England, 1989.

[8] B. F. Skinner, Behavior of Organisms, Appleton-Century-Crofts, 1938.

[9] D. S. Touretzky, L.M. Saksida, "Operant conditioning in skinnerbots.", Adaptive Behavior, 5(3/4), pp. 219-247, 1997.

[10] L. Kaelbling, M. Littman, A. Moore, "Reinforcement Learning : A Survey", J. Artificial Intelligence Research, vol. 4, pp. 237285, 1996.

[11] W. S. Lovejoy, A survey of algorithmic method for partially observable Markov decision processes., Annual of Operation Research, 28, pp47-66, 1991.

[12] R. Sun, C. Sessions, Self Segmentation of Sequences., IEEE Trans System Man and Cybernetics, vol. 30, no. 3, pp. 403418, 2000.

[13] M .Wieringm, J. Schmidhuber, "HQ-learnming. Adaptive Behavior", 6:2, pp. 219-246, 1997.

[14] M. Humphrys, "Action selection methods using reinforcement learning", From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior, Cambridge, MA, pp 135-144, MIT Press, 1996.

[15] L. Chrisman, "Reinforcement Learning with Perceptual Aliasing : The Perceptual Distinctions Approach", National Conference on Artificial Intelligence, pp 183-188, 1992.

[16] R. Sun and T. Peterson, "Autonomous Learning of Sequential Tasks: Experiments and Analyses", IEEE Trans. Neural Networks, vol.9, no. 6, Nov. 1998.

[17] R.E. Neapolitan, Foundation of algorithms : using C++ pseudocode, Jones and Bartlett Publishers, 1998.

저 자 소 개

권 우 영 (權 又 英)

2001년 한양대학교 공과대학 기계공학과 졸업. 2003년 한양대학교 정보통신대학원 졸업 정보통신공학과 졸업. 관심분야는 인공생명 및 기계학습

이 상 훈 (李 相 勳)

1994년 한양대학교 이과대학 수학과 (이학사). 1997년 한양대학교 산업대학원 전자계산학과(공학석사). 2000년 ~ 현재 한양대학교 전자전기제어계측과 박사과정 재학중. 관심분야 : 지능로봇의 행동선택 및 학습

서 일 홍 (徐 一 弘)

1977년 서울대학교 졸업. 1982년 한국과학기술원 졸업(공학박사). 1982년~1985년 대우 중공업 기술연구소 근무. 1987-1988년 미국 미시간대 객원 연구원. 1985년~현재 한양대학교 교수.