비선형 함수 근사화를 사용한 TD학습에 관한 연구

°권재철* 이영석** 김독옥* 서보혁***
*경북대학교 전기공학과 **영진전문대 전기과 ***경북대 전자전기공학부

A study of Temperal Difference Learning using Nonlinear Function Approximation

°Jae-cheol Kwon* · Young-seog Lee** · Dong-ok Kim* · Bo-hyeok Seo***
* Dept. of Electrical Eng. K.N.U. **Young-jin Junior college ***School of Electronic & Electrical Eng. K.N.U

Abstract - This paper deals with temporal-difference learning that is a method for approximating long-term future cost as a function of current state in knowlege-poor environment. a function approximator is used to approximate the mapping from state to future cost. a linear function approximator is limited because mapping from state to future cost has a nonlinear characteristic. so a nonlinear function approximator is used to approximate the mapping from state to future cost in this paper. and that TD learning using a nonlinear function approximator is stable is proved

1. 서 론

대부분의 교시 학습(supervised learning)과 비교시 학습(unsupervised learning)은 다양한 응용을 위하여 정확한 학습 자료를 필요로 한다. 하지만 상당한 실제 적용에서 정확한 학습 자료를 얻는 것은 어렵고 많은 대가를 치루어야 한다. 이러한 이유에서 강화 학습 알고리듬에 대한 관심이 커져왔다.[1]

강화 학습 문제에 대한 학습 자료는 대략적이고 단지 평가의 역할만 한다. 강화 학습에서는 강화 신호라고 부르는 스칼라 값인 평가 궤환만을 가지고 학습한다.

강화 학습은 목표를 이루기 위해 상호 작용으로부터 학습을 하는 방법으로 보상(reward)을 최대로 하도록 하는 상태과 행동과의 사상 관계를 학습하는 방법이다.

TD(Temporal-Difference) 학습[2]은 현재 상태의 함수로 미래의 보상를 근사화 하기 위한 방법이다.

선형 근사화 함수[3-4]를 사용하는 경우에는 상태에서 미래의 보상으로의 사상이 비선형성을 갖기 때문에 선형 근사화 함수를 이용하여 근사화하는데 한계가 있다.

본 논문에서는 환경 변수의 함수로 미래의 보상을 근사화하는데 비선형 근사화 함수로 신경망을 사용하고 학습률의 조정에 따른 리야프노프 관점에서의 수렴성을 보장하여 이를 사용한 TD학습이 안정함을 보임으로써 그 유용성을 보이고자 한다.

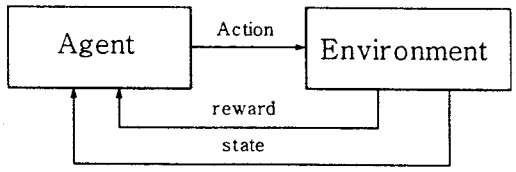

그림 1. 강화 학습의 구조

Fig. 2 The reinforement learning framework

2. 본 론

2.1 강화학습

동적으로 변화하는 환경에서는 환경을 한가지로 모델링할 수 없기 때문에 기존의 모델에 의존한 방법으로는 환경변화에 대처능력이 없게 된다. 강화 학습법은 자신과 환경과의 상호관계와 이에 따른 강화신호를 통하여 자신의 행동을 개선해 나가는 방법으로서 환경에 대한 정확한 사전지식이 없이 학습과 적응성을 보장하기 때문에 유용하다.

강화 학습은 목표를 이루기 위해 상호작용으로부터 학습을 하는 방법이다. 이때 학습자를 행위자(Agent)라고 하고 행위자를 제외한 모든 것을 환경(Environment)이라고 부른다. 행위자는 행동(Action)을 결정하고 환경(Environment)은 그 행동에 대해 응답(reward)을 하고 행위자에 새로운 상황(state)을 전달한다.

강화 학습은 스칼라 값인 보상치 혹은 강화 신호(reward)로 불리는 값을 최대로 하도록 하는 상황(state)과 행동(action)과의 사상 관계를 학습하는 방법이다. 행위자는 어떤 행동을 할 것인가를 배우는 것이 아니라. 어떤 행동이 가장 큰 보상을 받을 수 있는지를 알아내는 것이다.

강화 학습의 목적은 보상을 최대화 하는 것이다. 만약 시간 t 이후에 받은 보상들이 $r_{t+1}, r_{t+2}, r_{t+3}, \cdots$ 라고하면 이 값들을 최대화하는 가장 간단한 방법은 식 (1)과 같은 전체 보상을 최대화 하는 것이다.

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_T \qquad (1)$$

여기서 T는 상호 작용이 끝나는 최종 시간이다.

반면 행위자와 환경과의 상호 작용이 끝나지 않고 계속 될 때는 문제가 발생한다. 위에서 주어진 전체 보상값이 무한 수의 합의 형태가 되는 것이다.

그래서 전체 보상의 감쇄 보상을 도입한다. 이에따라 행위자는 각각의 시간에서 식 (2)와 같은 감쇄보상을 최대화하도록 학습하는 것이다.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \qquad (2)$$

여기서 γ는 감쇄 인자이다. 감쇄 인자는 미래의 보상들이 현재의 가치에 주는 영향의 정도를 결정한다.

감쇄 인자가 $0 \le \gamma < 1$ 이면 무한의 감쇄 보상은 보상들이 제한됨에따라 유한하게된다. $\gamma = 0$ 이면 행위자는 단지 즉시의 보상들을 최대화하는데만 관련된다. 이 경우 행위자는 단지 r_{t+1}을 최대화하기 위해 각 시간에서 어떻게 행동할 것인가를 학습하게 된다. γ가 1에 접근하면 미래의 보상을 더욱 중요시하게 된다.

2.2 TD 학습법 (Temperal Difference Learning)

TD 학습법은 Sutton이 제안한 강화 학습법으로 현재 상태의 함수로 미래의 보상를 근사화하는 방법이다. TD학습은 모델을 필요로 하지 않고 새로운 경험으로부터 직접적으로 배우게 되며, 마지막 결과가 나올 때 까지 기다리지 않고 학습된 예상치에 어느 정도 준하여 새로운 예상치를 얻는다.

식 (3)은 가장 간단한 TD 학습법(TD(0))이다.

$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)) \quad (3)$$

여기서 V는 전체 보상을 나타내는 가치함수, s_t는 t 스텝에서의 상태, r_{t+1}은 t+1 스텝에서의 보상치, α는 양의 학습률이다.

2.2.1 Actor-Critic Methods

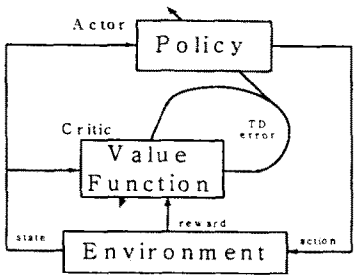

그림 2 Actor-Critic의 구조

Actor-Critic 방법[5]은 평가함수(value function)를 표현하는 부분과 정책함수(Policy)를 표현하는 부분으로 나눈 TD 학습법이다.

정책함수 부분은 행동(action)을 선택하는데 사용되기 때문에 Actor라고 하고, 가치함수 부분은 Actor에의한 행동을 평가하기 때문에 Critic이라고 한다. 평가값은 TD오차 형태로 취해지고 이 스칼라 신호는 Critic의 유일한 출력이고 Actor와 Critic의 학습에 이용된다. 새로운 행동이 선택된 후 Critic은 새로운 상태가 평균값보다 나쁜지 좋은지를 결정한다.

이 평가값이 식 (4)에 나타난 TD 오차이다.

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (4)$$

평가 함수는 각 주어진 상태의 평가값을 계산하며, 한 상태의 평가값은 미래에 기대되는 보상과 관련이 있으며 현 상태의 유용성을 포함하는 개념이다. 정책 함수는 현재의 상태로부터 다음 행위를 얻어내게되며, 평가함수와 정책함수 모두는 환경과 상호작용으로부터 학습된다.

얻어진 오차는 평가 함수와 정책 함수를 학습하는데 사용된다. 식 (5)과같이 각 스텝 t에 있어서 미래 보상값의 감쇄합을 예측 결과로 사용한다.

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \quad (5)$$

따라서 부정확한 예측에서 발생하는 차이를 TD오차라고 하며 이는 $r_{t+1} + \gamma V_{t+1} - V_t$로 주어진다. 바로 이 신호가 평가 함수 혹은 정책 함수의 오차 신호이며 결국 이 값을 줄이는 방향으로 학습이 이루어지게 된다.

2.3 함수 근사화 (Function Approximation)

함수 근사화는 상태로부터 미래의 보상으로의 사상을 근사화하는데 사용된다. 함수 근사화의 변수들은 상태변화와 그에 관련된 보상의 관측값에 의해 갱신된다. 함수 근사화의 목적은 더 많은 상태를 관측함에따라 장기적인 보상의 근사화 값을 향상시키는 것이다.

2.3.1 기울기 강하법 (Gradient-Descent Methods)

기울기 강하법을 기반으로하는 보상예측에서의 함수 근사화를 위한 학습방법을 살펴본다. 기울기 강하법은 모든 함수 근사화 방법에 널리 사용되고 특히 강화학습에 적합하다.

변수 벡터는 식 (6)와 같이 나타낼 수 있다.

$$\vec{\theta_k} = (\theta_k(1), \theta_k(2), \ldots, \theta_k(n))^T \quad (6)$$

상태가 동일한 분포로 나타난다고 가정한다.

MSE(Minimun Square Error)가 최소가 되도록 변수를 갱신한다.

$$\begin{aligned} \vec{\theta}_{k+1} &= \vec{\theta}_k - \frac{1}{2}\alpha \nabla_{\vec{\theta}_k}(V^*(s_t) - V_t(s_t))^2 \\ &= \vec{\theta}_k + \alpha(V^*(s_t) - V_t(s_t)) \nabla_{\vec{\theta}_k} V_t(s_t) \quad (7) \\ &= \vec{\theta}_k + \alpha(r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t)) \nabla_{\vec{\theta}_k} V_t(s_t) \end{aligned}$$

여기서 $\nabla_{\vec{\theta}_t} f(\vec{\theta}_t)$ 는 임의의 함수 f 에대해 변수 $\vec{\theta}_t$에대한 편미분이다.

2.4. 신경망 (Neural Network)

신경망은 기본적으로 학습을 통해 환경에 적응할 수 있는 능력과 뉴런들 사이의 연결강도를 조정하고, 이 연결 강도와 뉴런 자체의 비선형성의 조합으로 임의의 비선형 함수를 표현할 수 있다. 학습되지 않은 구역에서도 학습된 내용에 근거하여 근사화된 결과를 준다. 학습한 내용은 연결강도에 분산되어 저장되므로 신경회로의 일부가 손상된다 하더라도 학습된 내용은 크게 손상 받지 않는다.

2.4.1 Multi-Layer Perceptron

MLP는 가장 많이 알려져 있는 교시 학습 신경망 모델 중 하나로 일반적으로는 여러 개의 은닉층을 가지고 있지만, 출력층과 하나의 은닉층, 입력층으로 층분이 비선형 함수를 표현할 수 있다[6]

신경망 학습을 위한 목적함수는 식 (8)과 같이 정의한다.

$$E = \frac{1}{2}(O(k) - d(k))^2 \quad (8)$$

여기서 O(k)는 k step에서의 신경망 출력, d(k)는 k step에서의 기준신호이다.

신경망의 구조가 그림 (3)과 같을 때를 고려한다.

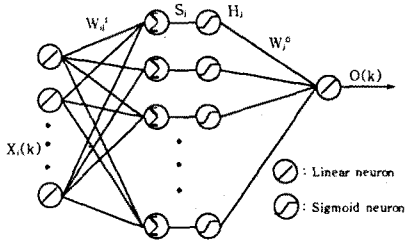

그림 3 다층 신경망

$$\begin{aligned} O(k) &= \sum_j W_j^o H_j(k) \\ H_j(k) &= f_a(S_j(k)) \quad (9) \\ S_j(k) &= \sum_i W_{ij}^I X_i(k) \end{aligned}$$

여기서 W_j^o는 출력층과 은닉층의 j번째 뉴런간의 가중치이고, W_{ij}^I는 은닉층의 j번째 뉴런과 입력층의 i번째

뉴런간의 가중치이고, $f_a(x) = \tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$를 활성화 함수로 사용한다.

가중치의 학습은 경사 학습법에의해 목적함수를 감소시키는 방향으로 갱신된다.

$$W(k+1) = W(k) + \eta(-\frac{\partial E}{\partial W}) \quad (10)$$
$$= W(k) + \Delta W(k+1)$$

여기서 가중치에 대한 목적함수의 변화량은 식 (11)와 같다.

$$\frac{\partial E}{\partial W} = \frac{\partial E}{\partial O(k)} \frac{\partial O(k)}{\partial W} \quad (11)$$
$$= e(k) \frac{\partial O(k)}{\partial W}$$

여기서 e(k) = O(k) - d(k) 이고, 각 가중치에 대한 신경망 출력의 변화량은 식 (12)과 식 (13)과 같다.

$$\frac{\partial O(k)}{\partial W_j^o} = H_j \quad (12)$$

$$\frac{\partial O(k)}{\partial W_{ij}^I} = W_j^o \, f_a(S_j(k)) X_i(k) \quad (13)$$

여기서 $f_a(\cdot)$는 활성화 함수를 미분한 값이다.

2.5 적응화 학습률을 이용한 수렴 보장

신경망의 수렴성은 학습률에 밀접한 관계가 있으며 적응화된 학습률은 수렴성 및 학습 성능에 효과적일 수 있다. 이와 같은 목적으로 다음과 같은 정리를 통해 함수 근사화에 사용된 신경망의 수렴성을 보장하는 학습률을 구한다.

정리 1 : 학습률의 초기치가 $0<\eta_1<2$ 이고 신경망의 학습률이 다음과 같이 갱신될 때 신경망의 수렴성이 점근적으로 보장된다.

$$\eta_k = \frac{\eta_1}{\|\nabla J\|^2} \quad (14)$$

여기서 η_1은 초기 학습률이다.

증명 : 함수 근사화에 사용된 신경망 가중치 갱신식을 식 (14)과 같이 표현하도록 하자

$$W_{k+1} = W_k + \eta_k \delta_k \nabla J_k \quad (15)$$

(15)식 양변에 최적의 가중치 W^*를 빼주면 (16)식과 같다.

$$W^* - W_{k+1} = W^* - (W_k + \eta_k \delta_k \nabla J_k) \quad (16)$$
$$\overline{W}_{k+1} = \overline{W}_k - \eta_k \delta_k \nabla J_k \quad (17)$$

여기서 $\overline{W}_k = W^* - W_k$이다.

그리고 식 (18)과 같은 비용함수를 고려한다.

$$V_k = \|\overline{W}_k\|^2 \quad (18)$$

비용함수의 변화치는 다음과 같다.

$$\Delta V_{k+1} = V_{k+1} - V_k \quad (19)$$
$$= \|\overline{W}_k\|^2 + \|\eta_k \delta_k \nabla J_k\|^2 - 2\eta_k \delta_k \nabla J_k \overline{W}_k$$
$$= \eta_k^2 \delta_k^2 \|\nabla J_k\|^2 - 2\eta_k \delta_k \nabla J_k \overline{W}_k$$
$$= -\eta_k(2\delta_k \nabla J_k \overline{W}_k - \eta_k \delta_k^2 \|\nabla J_k\|^2)$$

여기서 $\nabla J_k \overline{W}_k = \nabla J_k (W^* - W_k) = \delta_t'$ 일 때 ΔV_{k+1}은 다음과 같이 정리된다.

$$\Delta V_{k+1} = -\eta_k(2\frac{\delta_k'}{\delta_k} - \eta_k \|\nabla J_k\|^2)\delta_k^2 \quad (20)$$
$$\le -\eta_k(2 - \eta_k \|\nabla J_k\|^2)\delta_k^2$$

여기서 식 (16)를 대입하면 식 (20)은 다음과 같이 된다.

$$\Delta V_{k+1} = -\frac{\eta_1(2-\eta_1)}{\|\nabla J_k\|^2}\delta_k^2 \le 0 \quad (21)$$

여기서 $\eta_1(2-\eta_1)>0$.

따라서 $0<\eta_1<2$ 이고 (14)식을 만족할 때, $\Delta V_{k+1}<0$이 되어서 리야프노프 안정도 이론에 의해 (18) 식으로부터 $V_{k+1}>0$. (21)식에서 $\Delta V_{k+1} < 0$ 임을 보임으로써 신경망의 안정도가 보장된다.

3. 결 론

본 논문에서는 학습을 위한 정확한 입출력 자료를 얻기 어려운 경우에 대하여 환경과의 상호작용을 통해 단지 평가의 역할만 하는 강화 신호를 이용하여 보상을 최대로하는 행동을 결정함으로써 학습하는 경우에 대해 현재의 환경변수의 함수로 미래의 보상을 근사화하는 방법인 TD 학습에서 변수와 미래의 보상과의 사상을 근사화하는 근사화 함수로, 신경망을 이용하였고 학습률의 조정에 따른 리야프노프 관점에서의 수렴성을 보장하여 이를 사용한 TD학습이 안정함을 보임으로써 그 유용성을 보였다.

[참 고 문 헌]

[1] G. E. Hinton, "Connectionis learning procedures," Artificial Intelligence,vol. 40,no. 1,pp.143-150,1989.

[2] R. S. Sutton, "learning to predict by the methods of temporal differences," Machine Learning, vol. 3, pp. 9-44, 1988.

[3] P.D.Dayan, "The convergence of TD(λ) for general λ," Machine Learning, vol. 8, pp. 341-362, 1992

[4] J. N. Tsitsiklis and B. V. Roy, " An Analysis of Temporal-Difference Learning with funtion Approximation", IEEE Trans. on Automatic Control, vol.42, pp.674-690

[5] Barto,A.G. and Sutton,R.S. and Anderson,C.W. "Neuronlike elements that can solve difficult learning control problems", IEEE Trans. on Systems, Man, and Cybernetics, vol. SMC-13, no. 5 1983

[6] P. D. Wasserman, Adavanced methods in neural computing, Van Nostrand Reinhold, New York, 1993.