개미 집단 시스템에서 TD-오류를 이용한 강화학습 기법

이 승 관† · 정 태 충††

요 약

강화학습에서 temporal-credit 할당 문제 즉, 에이전트가 현재 상태에서 어떤 행동을 선택하여 상태전이를 하였을 때 에이전트가 선택한 행동에 대해 어떻게 보상(reward)할 것인가는 강화학습에서 중요한 과제라 할 수 있다. 본 논문에서는 조합최적화(hard combinational optimization) 문제를 해결하기 위한 새로운 메타 휴리스틱(meta heuristic) 방법으로, greedy search뿐만 아니라 긍정적 반응의 탐색을 사용한 모집단에 근거한 접근법으로 Traveling Salesman Problem(TSP)를 풀기 위해 제안된 Ant Colony System(ACS) Algorithms에 Q-학습을 적용한 기존의 Ant-Q 학습방법을 살펴보고 이 학습 기법에 다양화 전략을 통한 상태전이와 TD-오류를 적용한 학습방법인 Ant-TD 강화학습 방법을 제안한다. 제안한 강화학습은 기존의 ACS, Ant-Q학습보다 최적해에 더 빠르게 수렴할 수 있음을 실험을 통해 알 수 있었다.

A Reinforcement Learning Method using TD-Error in Ant Colony System

SeungGwan Lee† · TaeChoong Chung††

ABSTRACT

Reinforcement learning takes reward about selecting action when agent chooses some action and did state transition in present state. this can be the important subject in reinforcement learning as temporal-credit assignment problems. In this paper, by new meta heuristic method to solve hard combinational optimization problem, examine Ant-Q learning method that is proposed to solve Traveling Salesman Problem (TSP) to approach that is based for population that use positive feedback as well as greedy search. And, suggest Ant-TD reinfoecement learning method that apply state transition through diversification strategy to this method and TD-error. We can show through experiments that the reinforcement learning method proposed in this paper can find out an optimal solution faster than other reinforcement learning method like ACS and Ant-Q learning.

키워드 : 개미집단최적화(Ant Colony Optimization : ACO), 개미시스템(Ant System : AS), 개미집단시스템(Ant Colony System : ACS), 다양화(Diversification), 메타휴리스틱(Meta Heuristic), TD-오류(TD-error)

1. 서 론

강화학습(Reinforcement Learning)은 학습을 수행하는 에이전트가 동적환경에 대해 시행-착오(trial-anderror)에 의해 상호 작용하면서 학습을 수행하기 때문에 상호 작용에 의한 학습이라고도 한다. 에이전트는 학습을 수행하는 동안 주어진 환경에서 취할 수 있는 행동(action)을 시도하며, 그 행동에 대한 강화값을 받아 강화된다.

본 논문에서는 TSP 문제를 풀기 위해 Colorni, Dorigo 그리고 Maniezzo[1, 2]에 의해 제안된 새로운 메타 휴리스틱(Meta Heuristic) 방법인 강화학습 기반의 조합최적화 문제를 해결하는 Ant-Q[8, 9] 학습방법을 살펴보고 이 Ant-Q학습에 TD-오류를 적용한 학습방법[3-5, 13]인 새로운 Ant-TD 강화학습 기법을 제안한다.

† 준 회 원 : 경희대학교 대학원 전자계산공학과
†† 정 회 원 : 경희대학교 컴퓨터공학과 교수
논문접수 : 2003년 8월 8일, 심사완료 : 2003년 12월 13일

제안된 Ant-TD강화학습 방법은 다양화 전략을 통한 상태전이와 TD-오류를 이용하여 목표상태를 탐색하는 학습방법으로 기존의 ACS, Ant-Q학습보다 최적해에 더 빠르게 수렴하는 특징이 있다.

본 논문은 5장으로 구성되어 있으며 각 장의 주요내용은 다음과 같다. 2장에서는 ACS 알고리즘과 Ant-Q 강화학습 방법에 대하여 기술하고, 3장에서는 본 논문에서 제안한 Ant-TD강화 학습 방법과 특징에 대해 설명한다. 4장에서는 강화 학습의 학습 성능을 평가하기 위해 ACS, Ant-Q, Ant-TD학습 성능을 비교 분석한다. 그리고 5장에서는 결론과 향후 연구과제를 제시한다.

2. 조합 최적화를 위한 휴리스틱 접근법

2.1 Ant System(AS)

Ant System(AS)은 실제 개미들이 먹이에서 집까지 가장

짧은 경로를 찾는 능력을 모방한 메타 휴리스틱 탐색[1, 2, 7, 12, 15]으로 최근에는 강화학습(Reinforcement Learnig)의 특별한 한 분야로 소개되고 있다[8, 9].

이 방법은 에이전트라 불리는 개미들이 목적지를 향해 나아가는 동안 각 경로에 페로몬을 분비하고, 이후에 지나가는 에이전트들은 그 경로에 쌓여있는 페로몬(Pheromone) 정보를 이용해 다음 경로를 선택하는 원리를 휴리스틱 탐색에 적용시킨 시스템으로, 에이전트들의 행위를 살펴보면 다음과 같다. 먼저 각 에이전트들이 특정 경로을 선택해야되는 결정지점(decision point)에 도달하게 되면 그들은 최선의 선택(best choice)에 관한 어떤 정보도 가지고 있지 않기 때문에 무작위로 다음 경로를 선택하고 선택 후 지나간 길에 페로몬을 분비한다. 그 후 각 에이전트들이 다음으로 방문할 경로를 선택할 때는 각 경로에 쌓여있는 페로몬 양에 비례해 길을 선택하고 얼마정도의 시간이 경과하게 되면, 이 페로몬 양은 이후에 새로운 에이전트들이 경로를 선택할 시에 영향을 줄 정도로 각 경로에서 커다란 양의 차이를 보이게 된다. 이러한 과정들이 지나면 에이전트들은 각 경로에 있는 페로몬 양을 기반으로 서로 간의 정보 교환을 통해 최적의 경로를 찾아가고 이러한 에이전트들의 행동 양식을 그대로 적용한 알고리즘이 Ant System(AS)이다[6, 7, 10, 11, 14].

AS에서 노드(r)에 있는 에이전트 k가 노드(s)로 이동할 확률은 식 (1)로 표현하며, random proportional action choice rule(or state transition rule)으로 불린다.

$$p_k(r,s)=\begin{cases}\dfrac{[\tau(r,s)]\cdot[\eta(r,s)]^{\beta}}{\sum_{u\in J_k(r)}[\tau(r,u)]\cdot[\eta(r,u)]^{\beta}} & \text{if } s\in J_k(r)\\ 0 & otherwise\end{cases} \quad (1)$$

여기서 $\tau(r,s)$는 노드(r)과 노드(s)사이의 간선의 페로몬 양, $\eta=1/\delta$은 $\delta(r,s)$ (노드(r)과(s)의 거리)의 역수이고, $J_k(r)$은 노드(r)에 있는 에이전트 k가 방문할 수 있는 남아있는 노드들의 집합이다. 그리고 β는 페로몬과 간선 길이의 상대적인 중요도를 결정하는 파라미터(parameter)이다($\beta>0$).

AS에서 전역 갱신은 모든 경로가 완성된 후 경로를 구성한 모든 간선에 대해 갱신시키는데 그 방법은 다음 식 (2)와 같다.

$$\tau(r,s)\leftarrow(1-\alpha)\cdot\tau(r,s)+\sum_{k=1}^{m}\Delta\tau_k(r,s)$$
$$where\ \Delta\tau(r,s)=\begin{cases}(L_k)^{-1}, & \text{if}(r,s)\in tour\ done\ by\ ant\ k\\ 0, & otherwise\end{cases} \quad (2)$$

$0<\alpha<1$는 페로몬 지연 파라미터(Pheromone decay parameter), L_k는 현재까지의 경로길이, m은 에이전트 수이며, 경로 길이가 짧을수록 더 많은 페로몬 갱신이 발생하는데 이것은 강화학습과 유사한 방법이다.

그러나, 이 AS는 에이전트들이 짧은 경로가 있으면 그것만을 선택하고자 하는 성질로 인하여 국부 최적(Local Minima)에 빠질 확률이 높아지기 때문에, 이 문제를 쉽게 해결하기 위해 ACS 알고리즘이라는 방법이 새롭게 연구되었다.

2.2 Ant-Q Algorithm

Coloni, Dorigo and Mauiezzo가 제안한 Ant-Q 학습방법[8, 9]은 기존의 Ant System(AS)의 확장으로, Q-학습의 관점에서 재해석된 강화 학습법이다. 그러나, Ant-Q학습은 상태공간을 하나의 에이전트로 탐색하는 일반적인 Q-학습과는 달리 협력 에이전트들의 집합을 이용해 학습을 수행한다. 이러한 에이전트들은 서로 협력하여 AQ-값(Q-학습에서 Q-value와 유사)으로 표현되는 정보를 교환한다.

Ant-Q에서 노드(r)에 있는 에이전트k가 노드(u)로의 이동은 다음의 pseudo-random proportional action choice rule(or state transition rule)에 의해 수행된다.

$$s=\begin{cases}\arg\max_{u\in J_k(r)}\{[AQ(r,u)]^{\delta}[HE(r,u)]^{\beta}\} & \text{if } q\le q_0(exp\ loitation)\\ S & otherwise(exp\ loration)\end{cases} \quad (3)$$

$AQ(r,u)$는 Ant-Q값으로, 간선E(r,u)에 관계된 양의 값(positive value)이다. $AQ(r,u)$는 Ant-Q에서 Q-학습의 Q-값과 상응하는 값으로, 노드(r)에서 노드(u)로 이동하는 것이 얼마나 유용한지를 나타내는 것으로 이동함에 따라 그 값이 갱신된다. $HE(r,u)$는 노드(r)에서 노드(u)를 선택할 때 우수성을 평가하는 휴리스틱 값(heurisitc value)이다(TSP 문제에서는 노드(r)에서 노드(u)사이 거리의 역수). $J_k(r)$은 현재 노드(r)에서 에이전트 k가 방문할 수 있는 남아있는 노드들의 집합이다. 그리고, 초기 $AQ(r,u)=AQ_0=1/(average_length_of_edges\cdot n)$이다.

파라미터 δ와 β는 AQ-값과 휴리스틱 값의 상대적 중요도를 나타낸다. q는 [0, 1] 사이에 정규적으로 분포된 무작위 파라미터(random parameter)이고, q_0는 [0, 1] 사이의 값을 가지는 파라미터를 나타내며, S는 노드(r)에서 노드(s)를 선택할 때 식 (4)의 확률분포에 따라서 선택된 임의 변수이다.

$$p_k(r,s)=\begin{cases}\dfrac{[AQ(r,s)]^{\delta}\cdot[HE(r,s)]^{\beta}}{\sum_{u\in J_k(r)}[AQ(r,u)]^{\delta}\cdot[HE(r,u)]^{\beta}} & \text{if } s\in J_k(r)\\ 0 & otherwise\end{cases} \quad (4)$$

Ant-Q의 목표는 확률적으로 더 나은 목표값을 찾을 수 있는 AQ-값을 학습하는 것이다. AQ-값은 식 (5)에 의해

갱신된다.

$$AQ(r,s) \leftarrow (1-\alpha)\cdot AQ(r,s) + \alpha\cdot(\Delta AQ(r,s) + \gamma\cdot \underset{z\in J_k(s)}{Max}\, AQ(s,z)) \quad (5)$$

$\alpha(0<\alpha<1)$는 페로몬 지연 파라미터로 학습율(learning rate), γ는 할인율(discount rate), $MaxAQ(s,z)$는 다음 상태에 대한 평가로 외부 환경으로부터 받는 강화값을 최대화하는 것으로 전역 강화일 때는 0이다.

또한, ΔAQ는 강화값으로 지역 강화일 때는 항상 0이다. 전역강화는 에이전트들이 모든 경로(tour)를 완성 후에 수행되는데 다음의 식 (6)에 의해 갱신된다.

$$\Delta AQ(r,s) = \begin{cases} \dfrac{W}{L_{kib}}, & \text{if } (r,s) \in \textit{tour done by the agent } k_{ib} \\ 0, & \textit{otherwise} \end{cases} \quad (6)$$

여기서, W는 상수값으로 여러 실험을 통해 $W=10$으로 고정한다. L_{kib}는 현재 경로 사이클의 최적 경로 길이(Current Iteration Best Tour Length)이다.

3. Ant-TD를 이용한 강화학습

강화학습에서 temporal-credit 할당 문제 즉, 에이전트가 현재 상태에서 어떤 행동을 선택하여 상태전이를 하였을 때 에이전트가 선택한 행동에 대해 어떻게 보상(reward)할 것인가는 강화학습에서 중요한 과제라 할 수 있다.

본 논문에서는 Ant-Q 강화 학습에서 Temporal-credit 할당 문제를 해결하기 위해 제시된 TD-오류를 이용한 강화학습방법을 제안한다.

제안된 Ant-TD 학습은 기존의 Ant-Q 학습 성능을 개선하기 위해 새롭게 제안된 방법으로 Ant-Q 학습에 C.J.C.H. Watkins가 제안한 TD-오류를 적용한 방법이다[3].

3.1 다양화를 통한 상태전이

Ant-Q에서 노드(r)에 있는 에이전트 k가 노드(s)로 이동할 확률은 식 (3)의 상태전이 규칙(state transition rule)에 의해 수행되는데 여기서 탐험(exploration)과 탐색(exploitation)의 균형은 매우 중요하다. 앞서 q와 q_0에 대해 기술하였는데, 이것은 새로운 간선에 대한 다양한 탐험(다양성)과 축적된 정보에 대한 탐색(강화) 사이에서의 균형을 직접적으로 유지하기 위하여 이용된다.

에이전트들은 경로를 탐색동안 휴리스틱 정보와 페로몬 정보를 이용하는데, 만약 $q<q_0$인 경우 에이전트는 탐색 행동을 취하는데 그것은 오랜 기간동안 최선의 선택으로 페로몬 정보와 짧은 기간동안 경험적 지식으로서 거리와 관련된 휴리스틱 값을 사용한다. $q>q_0$인 경우에는 에이전트는 다양한 탐험를 수행하기 위하여 확률 분포(S)를 사용하여 길이가 더 짧고 많은 양의 페로몬을 가진 간선 선택을 선호하게 된다. 그러나, 식 (4)를 적용한 Ant-Q에서는 최적 경로에 속한 간선 선택 확률만 높게 만들기 때문에 에이전트들은 확률 분포(S) 적용의 기본 목적인 다양한 탐험 수행을 할 수 없게 된다.

따라서, 현재 최적 경로로 새로운 이웃 간선으로의 다양한 탐험의 역할을 고려해 각 간선에 에이전트들의 방문 횟수를 고려할 수 있는데, 이것은 탐험의 비율을 빨리하고 탐험의 정확성을 점차로 개선함으로써 강화학습에 있어서 효과적일 수 있다.

다양한 탐험의 역할을 강화하기 위해서 확률분포로 방문 횟수를 적용할 수 있는데, 방문 횟수를 다양한 탐험에 적용하기 위해 식 (4)를 다음 식 (7)과 같이 수정할 수 있다. 여기서 파라미터 δ는 현재 경로 사이클 동안 각 에이전트들이 간선 $E(r,s)$에 방문한 횟수를 확률적으로 적용한 것이다.

$$p_k(r,s) = \begin{cases} \dfrac{[\tau(r,s)]^{\delta}\cdot[\eta(r,s)]^{\beta}}{\sum_{u\in J_k(r)}[\tau(r,u)]^{\delta}\cdot[\eta(r,u)]^{\beta}} & \text{if } s\in J_k(r) \\ 0 & \textit{otherwise} \end{cases} \quad (7)$$

$$\textit{where, if } f(r,s)\neq 0 \quad \delta = -1/f(r,s)$$
$$\textit{else} \qquad\qquad \delta = 1$$

$f(r,s)$는 현재까지 에이전트들이 간선 $E(r,s)$를 방문한 빈도수로, 에이전트들이 해당 간선에 대해 한번이라도 방문했을 경우 δ는 방문 빈도수에 대한 음의 역수(-1/방문 빈도수)를 취한다. 그러나, 현재 경로 사이클에서 해당 간선에 대해 한번도 방문하지 않은 경우 $\delta=1$로 초기화한다. 결국 식 (7)에 의해 에이전트들은 새로운 탐색 공간으로의 다양한 탐색으로 인해 더욱 다양하게 새로운 간선을 선택할 수 있게 된다. 즉, 학습 에이전트들은 더 많은 페로몬을 가진 상태 대신에 방문 빈도수가 적은 상태를 선택 할 수 있게 됨으로써, 임의의 적합한 정책에 대한 집중을 회피하고 최적해를 신속히 탐색하게 하는 장점이 있다.

3.2 TD오류를 이용한 강화

Temporal-credit 할당 문제를 해결하기 위해 제시된 TD를 이용한 강화 학습은 예측을 학습하기 위한 방법이라고 할 수 있다. 예측을 학습하기 위해 교사 학습과 같은 전통적인 학습 방법들은 중간 단계에서 계산된 모든 출력에 대한 예측들에 대한 기록을 유지하면서 최종 결과가 발생할 때까지 기다린다. 그리고 나서 훈련 오류로서 최종 결과와 각 상태의 출력에 대한 예측 값과의 차를 이용한다. 그러나 TD를 이용한 강화 학습은 최종 결과를 기다리지 않고 학습한다. 매 학습 단계에서 훈련 오류는 현재 상태의 출력에 대한 예측과 다음 상태의 출력에 대한 예측과의 차를 이용

하며 이를 TD-오류(Temporal Difference error)라 한다. TD-오류를 이용한 TD-학습에서 현재 상태는 현재 상태의 출력에 대한 예측은 다음 상태의 출력에 대한 예측과 가깝게 하기 위해 갱신된다. 이와 같이 각 상태의 출력에 대한 예측 값들의 차이를 이용하는 단순한 형태의 강화 학습을 TD(0)-학습이라고 한다.

TD-오류를 이용한 TD(0)-학습은 현재 상태에 대한 다음 상태에 대한 예측과의 차이를 이용하여 현재 상태의 Q-함수 값을 식 (8)과 같이 계산한다.

$$Q(s_t, a_t) \leftarrow (1-\alpha) \cdot Q(s_t, a_t) + \alpha \cdot TDerror \quad (8)$$

여기서, α는 학습율(learning rate), r_t는 강화 값, γ는 할인율(discount rate),TD error는 현재 상태에 대한 다음 상태에 대한 예측과의 차이(TD-오류)로서 식 (9)와 같이 계산한다.

$$TD\,error = r_{t+1} + \gamma [\underset{a \in A(s_t)}{Max}\, Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad (9)$$

위의 TD-오류를 Ant-TD 강화학습에 적용하면, 현재 상태, 즉 노드(r)에서 선택된 노드(s)에 대한 Q-값($AQ(r,s)$)과 현재 상태의 노드(r,s) 쌍에 의해 선택된 다음 상태의 노드(s,z) 중에서 최대 Q-값($MaxAQ(s,z)$)을 갖는 노드(s,z) 쌍과의 Q-함수 값을 갱신하기 위해 TD-오류를 이용한다. TD-오류는 식 (10)과 같이 계산된다.

$$TD\,error = \Delta AQ(r,s) + \gamma [\underset{z \in J_k(s)}{Max}\, AQ(s,z) - AQ(r,s)] \quad (10)$$

결국, Ant-TD 강화학습은 TD-오류를 이용하여 최적의 값-함수를 구하기 위해 현재 상태의 노드(r,s) 쌍에 대한 Q-함수 값을 식 (11)과 같이 계산한다.

$$AQ(r,s) \leftarrow (1-\alpha) \cdot AQ(r,s) + \alpha(\Delta AQ(r,s) + \gamma [\underset{z \in J_k(r)}{Max}\, AQ(s,z) - AQ(r,s)]) \quad (11)$$

$\alpha(0 < \alpha < 1)$는 페로몬 지연 파라미터로 학습율(learning rate), γ는 할인율(discount rate), $MaxAQ(s,z)$는 다음 상태에 대한 평가로 외부 환경으로부터 받는 강화값을 최대화하는 것으로 전역 강화일 때는 0이다.

또한, ΔAQ는 강화값으로 지역 강화일 때는 항상 0이다. 전역강화는 에이전트들이 모든 경로(tour)를 완성 후에 수행되는데 다음의 식 (12)에 의해 갱신된다.

$$\Delta AQ(r,s) = \begin{cases} \dfrac{W}{L_{kib}}, & \text{if } (r,s) \in \textit{tour done by the agent } k_{ib} \\ 0, & \textit{otherwise} \end{cases} \quad (12)$$

여기서, W는 상수값으로 여러 실험을 통해 $W=10$으로 고정한다. L_{kib}는 현재 경로 사이클의 최적 경로 길이이다. Ant-TD 알고리즘은 (그림 1)과 같다.

```
/* Initialization phase */
   Set an initial value for AQ-values
/* Main algorithm */
   Loop /* This loop is an iteration of the algorithm */
      1. /* Initialization of agents data structures */
         Choose a starting node for agents
      2. /* In this step agents build tours and locally update
         AQ-values */
            Each agent applies the state transition rule(3) to choose
            the node to go to, updates the set J_k and applies Eq.(11)
            to locally update AQ-values (in Eq.(11) ΔAQ(r, s) = 0)
      3. /* In this step agents globally update AQ-values */
            The edges belonging to the tour done by the best agent
            are updated using Eq.(11) where ΔAQ(r, s) is given by
            Eq.(12)
   Until (End_condition = True)
```

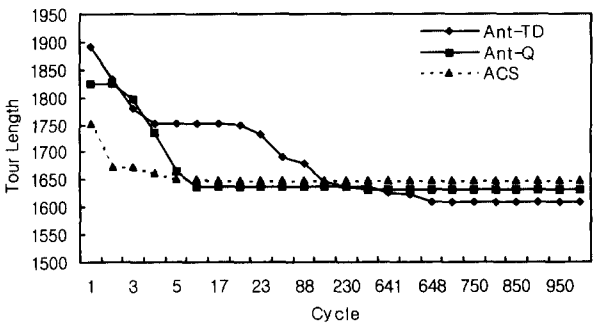
(그림 1) The Ant-TD Algorithm

4. 성능 측정 및 분석

본 논문에서 제안하고 있는 방법을 실험하기 위해서 사용된 파라미터들의 값은 실험에 의해 다음과 같이 결정되었으며, 일반적으로 실험에 의해 결정된 최적의 값은 δ= 식 (7), $\beta=2$, $\alpha=0.1$, $q_0=0.9$, $\gamma=0.3(0.2\sim0.6)$, $\lambda=0.3$, $m=n$, $W=10$, $AQ_0=1/(\text{average_length_of_edges} \cdot n)$이다.

ACS, Ant-Q 학습과 본 논문에서 제안한 Ant-TD 강화학습과의 효율성 분석을 위해 최적해에 얼마나 빨리 수렴하는가를 임의의 도시 집합들에 대해 비교 분석 하였다.

(그림 2)는 bayg29.tsp를 이용해 1000번 사이클을 반복했을 경우의 수렴속도를 보여주는 것이다. 초기에는 Ant-Q가 수렴속도가 빠르지만 탐색과정을 진행함에 따라 Ant-TD가 빠르게 수렴함을 볼 수 있다. 이것은 탐색 초기에 다양화 전략에 의한 상태전이 탐색 과정을 비탐사 지역으로 유도하기 위해 빈도수에 기반한 조건을 적용하여 사용함으로써 초기 수렴속도는 느리지만 탐색을 진행함에 따라 신속히 최적해를 탐색하게 한다. 결국, 다양한 상태전이와 TD-오류를 이용한 강화가 효과적임을 보여주는 것이다.



(그림 2) 수렴속도 평가

<표 1>은 (R×R 격자 문제)에 대해 10회 시행에 1000번 사이클을 반복했을 경우, 각 알고리즘에 의해 산출된 최적 경로 길이와 평균 경로 길이를 보여주는 것으로 문제 영역이 작은 경우, 즉 노드수가 작은 경우에는 결과의 차이가 없었지만 문제 영역이 커질수록 제안된 방법의 성능이 우수하다는 것을 보여주고 있다. 따라서 문제영역이 큰 문제에 대해 효과적으로 적용될 수 있다.

<표 1> Ant-TD 성능평가

Node	ACS		Ant-Q		Ant-TD	
	Average length	Best length	Average length	Best length	Average length	Best length
4×4	**160**	**160**	**160**	**160**	**160**	**160**
5×5	**254.14**	**254.14**	**254.14**	**254.14**	**254.14**	**254.14**
6×6	**360**	**360**	361.66	**360**	**360**	**360**
7×7	510.32	506.50	509.38	502.43	**494.14**	**494.14**
8×8	660.54	654.14	653.61	648.28	**640**	**640**

5. 결론과 앞으로의 연구방향

본 논문에서는 Ant-Q 강화 학습에서 Temporal-credit 할당 문제를 해결하기 위한 TD-오류 이용과 각 노드에 대해 에이전트들의 방문 빈도수 기반의 다양화 전략에 의한 상태전이를 통해 강화학습하는 방법을 제안하였다.

제안된 Ant-TD 강화 학습법은 기존의 Ant-Q 학습 성능을 개선하기 위해 새롭게 제안된 방법으로 에이전트들이 경로 사이클을 이루는 동안 방문한 간선에 대한 방문 빈도수를 상태전이 규칙에 적용해 에이전트들이 탐색영역을 더욱 다양하게 검색하게 하여 아직까지 탐색하지 않은 새로운 영역으로 다양하게 찾아가게 하고, TD-오류를 이용해 Temporal-Credit할당 문제를 해결하게 함으로써 최적해에 빠르게 수렴하게 하였다.

향후 연구과제는 Ant-TD에서 현재 상태에서 선택한 노드에 대해 얼마나 적합한가를 의미하는 척도인 적합도(Eligibility factor)를 이용한 강화학습 방법에 대한 연구도 있어야 겠다.

참 고 문 헌

[1] A. Colorni, M. Dorigo and V. Maniezzo, "An investigation of some properties of an ant algorithm," Proceediings of the Parallel Parallel Problem Solving from Nature Conference (PPSn'92), R. Manner and B. Manderick (Eds.), Elsevier Publishing, pp.509-520, 1992.

[2] A. Colorni, M. Dorigo and V. Maniezzo, "Distributed optimization by ant colonies," Proceedings of ECAL'91 - European Conference of Artificial Life, Paris, France, F. Varela and P. Bourgine (Eds.), Elsevier Publishing, pp. 134-144, 1991.

[3] C. J. C. H. Watkins, "Learning from Delayed Rewards," Ph. D. thesis, King's College, Cambridge, U.K, 1989.

[4] C. N. Fiecher, "Efficient reinforcement learning," In Proceedings of the Seventh Annual ACM Conference On Computational Learning Theory, pp.88-97, 1994.

[5] E. Barnald, "Temporal-difference methods and markov model," IEEE Transactions on Systems, Man, and Cybernetics, 23, pp.357-365, 1993.

[6]. L. M. Gambardella and M. Dorigo, "Solving symmetric and asymmetric TSPs by ant colonies," Proceedings of IEEE International Conference of Evolutionary Computation, IEEE-EC'96, IEEE Press, pp.622-627, 1996.

[7] L. M. Gambardella and M. Dorigo, "Ant Colony System : A Cooperative Learning approach to the Traveling Salesman Problem," IEEE Transactions on Evolutionary Computation, Vol.1, No.1, 1997.

[8] L. M. Gambardella and M. Dorigo, "Ant-Q : a reinforcement learning approach to the traveling salesman problem," Proceedings of ML-95, Twelfth International Conference on Machine Learning, A. Prieditis and S. Russell (Eds.), Morgan Kaufmann, pp.252-260, 1995.

[9] M. Dorigo and L. M. Gambardella, "A study of some properties of Ant-Q," Proceedings of PPSN IVFourth International Conference on Parallel Problem Solving From Nature, H.M.Voigt, W. Ebeling, I. Rechenberg and H.S. Schwefel (Eds.), Springer-Verlag, Berlin, pp.656-665, 1996.

[10] M. Drigo, V. Maniezzo and A. Colorni, "The ant system : optimization by a colony of cooperation agents," IEEE Transactions of Systems, Man, and Cybernetics-Part B, Vol.26, No.2, pp.29-41, 1996.

[11] M. Dorigo and G. D. Caro, "Ant Algorithms for Discrete Optimization," Artificial Life, Vol.5, No.3, pp.137-172, 1999.

[12] M. Dorigo and L. M. Gambardella, "Ant Colonies for the Traveling Salesman Problem," *BioSystems*, 43, pp.73-81, 1997

[13] R. C.Yee, P. E. Utgoff and A. G. Barto, "Explaining temporal differences to create useful concepts for evaluating states," In Proceedings of the 8th National Conference on Artificial Intelligence, pp.882-888, 1990.

[14] T. Stutzle and H. Hoos, "The ant system and local search for the traveling salesman problem," Proceedings of ICEC 1997~1997 IEEE 4th International Conference of Evolutionary.

[15] T. Sttzle and M. Dorigo, "ACO Algorithms for the Traveling Salesman Problem," In K. Miettinen, M. Makela, P. Neittaanmaki, J. Periaux, editors, Evolutionary Algorithms in Engineering and Computer Science, Wiley, 1999.

이 승 관

e-mail : lee@iislab.kyunghee.ac.kr

1997년 경희대학교 전자계산공학과 (공학사)

1999년 경희대학교 대학원 전자계산공학과 (공학석사)

2004년 경희대학교 대학원 전자계산공학과 (공학박사)

관심분야 : 인공지능, 지능에이전트, 메타알고리즘, 스케줄링

정 태 충

e-mail : tcchung@iislab.kyunghee.ac.kr

1980년 서울대학교 전자공학과(공학사)

1982년 한국과학기술원 대학원 전자계산공학과(공학석사)

1987년 한국과학기술원 대학원 전자계산공학과(공학박사)

1987년~1988년 KIST 시스템 공학센터 선임 연구원

2001년 미국 Iowa대학 교환교수

1988년~현재 경희대학교 컴퓨터공학과 정교수

관심분야 : 인공지능, 지능에이전트, 메타알고리즘