Temporal Difference 학습을 이용한 다중 집단 강화·다양화 상호작용 개미 강화학습

Multi Colony Intensification·Diversification Interaction Ant Reinforcement Learning Using Temporal Difference Learning

이승관
가톨릭대학교 컴퓨터정보공학부
Seung-Gwan Lee(leesg@catholic.ac.kr)

요약

본 논문에서는 Temporal Difference 학습을 적용한 Ant-Q 기반 개미 모델을 이용한 다중 집단 상호작용 개미 강화학습 모델을 제안한다. 이 모델은 몇 개의 독립적 개미시스템 집단으로 이루어져 있으며, 상호작용은 집단간 엘리트 전략(강화, 다양화 전략)에 따라 임무를 수행한다. 강화 전략은 다른 에이전트 집단의 휴리스틱 정보를 이용해 좋은 경로 선택을 가능하게 한다. 이것은 집단간 긍정적 상호작용을 통해 에이전트들의 방문 빈도가 높은 간선을 선택하게 한다. 다양화 전략은 에이전트들이 다른 에이전트 집단의 탐색 정보에 의해 부정적 상호작용을 수행함으로써 방문 빈도수가 높은 간선의 선택을 회피하게 만든다. 이러한 전략을 통해 제안한 강화학습은 기존의 개미집단시스템, Ant-Q학습보다 최적해에 더 빠르게 수렴할 수 있음을 실험을 통해 알 수 있었다.

■ 중심어 : | 강화학습 | 다중 집단 개미모델 | 상호작용 |

Abstract

In this paper, we suggest multi colony interaction ant reinforcement learning model. This method is a hybrid of multi colony interaction by elite strategy and reinforcement learning applying Temporal Difference(TD) learning to Ant-Q learning. Proposed model is consisted of some independent AS colonies, and interaction achieves search according to elite strategy(Intensification, Diversification strategy) between the colonies. Intensification strategy enables to select of good path to use heuristic information of other agent colony. This makes to select the high frequency of the visit of a edge by agents through positive interaction of between the colonies. Diversification strategy makes to escape selection of the high frequency of the visit of a edge by agents achieve negative interaction by search information of other agent colony. Through this strategies, we could know that proposed reinforcement learning method converges faster to optimal solution than original ACS and Ant-Q.

■ keyword : | Reinforcement Learning | Multi Colony Ant Model | Interaction |

접수번호 : #050221-001
접수일자 : 2005년 2월 21일
심사완료일 : 2005년 07월 12일
교신저자 : 이승관, e-mail : leesg@catholic.ac.kr

I. 서 론

최근 개미 모델[1-3]은 강화학습(Reinforcement Learning)의 특별한 한 분야로 소개되고 있다[4][5].

강화학습에서 에이전트가 현재 상태에서 어떤 행동을 선택하여 상태전이를 하였을 때, 에이전트가 선택한 행동에 대해 어떻게 보상할 것인가는 가장 중요한 과제라 할 수 있다.

본 논문에서는 Temporal Difference(TD) 학습을 적용한 Ant-Q 기반 개미 모델을 이용한 다중 집단 상호작용 개미 강화학습 모델(Multi Colony Interaction Ant Reinforcement Learning Model)을 제안한다.

TD 학습을 이용한 Ant-Q 기반의 다중 집단 상호작용 개미 모델은 매 학습 단계에서 현재 상태의 출력에 대한 예측과 다음 상태의 출력에 대한 예측과의 차이를 이용하여 학습한다[6][7]. TD 학습은 현재 상태에서 현재 상태의 출력에 대한 예측은 다음 상태의 출력에 대한 예측과 가깝게 하기 위해 갱신된다.

본 논문에서 제안된 개미 모델은 기존의 Ant-Q 모델 성능을 개선하기 위해 새롭게 제안된 방법이다. 이 방법은 Ant-Q 모델에 TD 학습을 적용한 강화학습과 엘리트 전략에 의한 다중 집단 상호작용 개미 모델을 적용한 혼합된 학습방법이다.

II. 기존 연구

2.1 Ant System

개미 시스템(Ant System : AS)은 실제 개미들이 먹이에서 집까지 가장 짧은 경로를 찾는 능력을 모방한 메타 휴리스틱 탐색[1-3]으로 최근에는 강화학습의 특별한 한 분야로 소개되고 있다[4][5].

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 경로를 선택해야 되는 결정지점에 도달하게 되면 그들은 최선의 선택에 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 알고리즘이다.

AS에서 노드(r)에 있는 에이전트(k)가 노드(s)로 이동할 확률은 식(1)로 표현하며, 상태전이 규칙(state transition rule)으로 불린다.

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

여기서 $\tau(r,u)$는 노드(r)과 노드(u)사이 간선의 페로몬의 양, $\eta(r,u)=1/\delta(r,u)$로써 $\delta(r,u)$는 노드(r)과 노드(u)사이 거리이며, $J_k(r)$은 노드(r)에 있는 에이전트(k)가 방문할 수 있는 남아있는 노드들의 집합이다. 그리고 β는 페로몬과 간선 길이의 상대적인 중요도를 결정하는 파라미터이다($\beta>0$).

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

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

$\rho(0<\rho<1)$는 페로몬 지연 파라미터(Pheromone decay

parameter), $\Delta\tau(r,s)$는 간선$E(r,s)$에 대해 전체 페로몬 증가량, $\Delta\tau_k(r,s)$는 에이전트(k)에 의한 간선$E(r,s)$의 페로몬 증가량, Q는 상수, L_k는 에이전트(k)의 경로길이, m은 에이전트 수이다. 여기서 경로 길이가 짧을수록 더 많은 페로몬 갱신이 발생하는데 이것은 강화학습 기법과 유사한 방법이다[4][5].

그러나 이 AS는 에이전트들이 짧은 경로가 있으면 그것만을 선택하고자 하는 성질로 인하여 국부 최적(Local Minima)에 빠질 확률이 높아지는 단점이 있다.

2.2 Ant-Q

Dorigo 등에 의해 제안된 Ant-Q 학습방법[4][5].은 기존의 AS의 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 에이전트(k)가 노드(u)로의 이동은 다음 식(3) 상태전이 규칙에 의해 수행된다.

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

$AQ(r,u)$는 Ant-Q값으로, 간선$E(r,u)$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. 초기 $AQ(r,u)=AQ_0=1/$(평균간선길이 $\bullet n$)이다.

$HE(r,u)$는 노드(r)에서 노드(u)를 선택할 때 우수성을 평가하는 휴리스틱 값(heurisitc value)이다(TSP문제에서는 노드(r)에서 노드(u)사이 거리의 역수).

$J_k(r)$은 현재 노드(r)에서 에이전트(k)가 방문할 수 있는 남아있는 노드들의 집합이며, 파라미터δ 와 β는 AQ-값과 휴리스틱 값의 상대적 중요도를 나타낸다.

q 는 [0,1] 사이에 정규적으로 분포된 무작위 파라미터이고, q_0는 [0,1]사이의 값을 가지는 파라미터를 나타낸다. S는 노드(r)에서 노드(s)를 선택할 때 식(4)의 확률분포에 따라서 선택된 임의 변수이다.

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

Ant-Q의 목표는 확률적으로 더 나은 목표값을 찾을 수 있는 AQ-값을 학습하는 것이다. AQ-값은 식(5)의 학습에 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)이며, γ 는 할인율(discount rate)이다. $MaxAQ(s,z)$는 다음 상태에 대한 평가로 외부 환경으로부터 받는 강화값을 최대화하는 것으로 전역 강화일 때는 0이다. 또한, $AQ(r,s)$는 강화값으로 지역 강화일 때는 항상 0이다.

전역 강화는 에이전트들이 모든 경로(tour)를 완성 후에 수행되는데 다음 식에 의해 갱신된다.

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

W는 상수값으로 여러 실험을 통해 $W=10$으로 고정한다. L_{kib}는 현재 경로 사이클의 최적 경로 길이이다.

2.3 Ant-TD

TD-학습은 현재 상태에 대한 예측과 다음 상태에 대한 예측과의 차이를 이용하여 현재 상태의 Q-함수 값을 식(7)과 같이 계산한다.

$$Q(s_t,a_t)\leftarrow(1-\alpha)\cdot Q(s_t,a_t)+\alpha\cdot TDerror \quad (7)$$

위 식에서, a는 학습율(learning rate)이며, *TDerror*는 현재 상태에 대한 예측과 다음 상태에 대한 예측과의 차이로써 식(8)과 같이 계산한다.

$$TD\ error = r_{t+1} + \gamma \cdot [\underset{a \in A(s_t)}{Max}\, Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad (8)$$

위 식에서, r_t는 강화 값, γ 는 할인율(discount rate)이다. 위의 TD-학습을 Ant-Q 개미 모델에 적용하면 식(9)와 같다[11]. 현재 상태의 노드(r)에 있는 에이전트(k)에 의해 선택된 노드(s)에 대한 Q-값($AQ(r,s)$)과 현재 상태의 노드(r,s)에 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 학습을 이용한다. TD 학습은 식(9)와 같이 계산된다.

$$TD\ error = \Delta AQ(r,s) + \gamma \cdot [\underset{z \in J_k(s)}{Max}\, AQ(s,z) - AQ(r,s)] \quad (9)$$

결국, Ant-Q 개미 모델에 TD 학습을 적용한 개미 강화학습 모델은 최적의 값-함수를 구하기 위해 현재 상태의 노드(r,s) 쌍에 대한 Q-함수 값을 식(10)과 같이 계산한다.

$$\begin{aligned} AQ(r,s) \leftarrow & (1-\alpha) \cdot AQ(r,s) \\ & + \alpha \cdot (\Delta AQ(r,s) + \gamma \cdot [\underset{z \in J_k(s)}{MaxAQ(s,z)} - AQ(r,s)]) \end{aligned}$$

$$\text{where} \quad \Delta AQ(r,s) = 0 \quad ,\textit{if Local updating}$$

$$\underset{z \in J_k(s)}{MaxAQ(s,z)} - AQ(r,s) = 0 \quad ,\textit{if Global updating} \quad (10)$$

$a(0<a<1)$는 학습율, γ 는 할인율이다. $MaxAQ(s,z)$는 다음 상태에 대한 평가로 외부 환경으로부터 받는 강화값을 최대화하는 것으로 전역 강화일 때는 0이며, $AQ(r,s)$는 강화값으로 지역 강화일 때는 항상 0이다.

III. TD 학습을 이용한 다중 집단 개미 강화학습

본 논문에서는 TD 학습을 적용한 Ant-Q 기반의 다중 집단 상호작용 개미 강화학습 모델을 제안한다.

제안된 개미 모델 구조는 몇 개의 독립적 AS 집단으로 이루어져 있으며, 상호작용은 강화 전략과 다양화 전략으로 나누어진 집단간 엘리트 전략에 따라 임무를 수행한다. 강화 전략은 다른 에이전트 집단의 휴리스틱 정보를 이용해 좋은 경로 선택을 가능하게 한다. 이것은 집단간 긍정적 상호작용을 통해 에이전트들의 방문 빈도가 높은 간선을 선택하게 한다. 다양화 전략은 에이전트들이 다른 에이전트 집단의 탐색 정보에 의해 부정적 상호작용을 수행함으로써 방문 빈도수가 높은 간선의 선택을 회피하게 만든다.

[그림 1]은 다중 집단 상호작용 개미 모델에서 집단간 엘리트 전략에 의한 상호연결망 구조를 보여주고 있다. 이 구조에서 상호작용은 집단간 엘리트 전략에 따라 페로몬 정보를 교환함으로써 이루어진다. [그림1.A]는 Mesh구조를 나타내는 것으로 문제 영역이 작은 경우에 효과적으로 적용 될 수 있는 구조이며, [그림1.B]는 Double Mesh구조로 문제 영역이 복잡하고 큰 경우에 효과적으로 적용 될 수 있는 구조이다

[그림1.A]는 유방향 연결망 구조로 되어 있으며, 집단(C1,C2), (C2,C3), (C3,C4), (C4,C1) 사이에는 다양화 엘리트 전략에 의한 부정적 상호작용을, 집단(C1,C5), (C2,C5), (C3,C5), (C4,C5) 사이에는 강화 엘리트 전략에 의한 긍정적 상호작용을 수행한다. [그림1.B]는 유방향 이중 연결망 구조로 되어 있으며, Queen1그룹에서 집단(C1,C5), (C2,C5), (C3,C5), (C4,C5) 사이에는 강화 전략에 의한 긍정적 상호작용을, 그 외의 집단 사이에서는 다양화 전략에 의한 부정적 상호작용을, Queen2그룹에서 집단(C6,C10), (C7,C10), (C8,C10), (C9,C10) 사이에는 긍정적 상호작용을, 그 외의 집단 사이에서는 부정적 상호작용을 한다. 그리고 C5, C10은 중심 에이전트 집단(여왕 집단, Queen Colony)이고 나머지 집단은 일개미 집단이다. 그리고 중심 집단(C5,C10)사이에서는 지역 최적해 상호 교환을 통한 정보 갱신이 수행된다. 이것은 각 집단사이에 서로 다른 임무를 수행함을 의미한다.

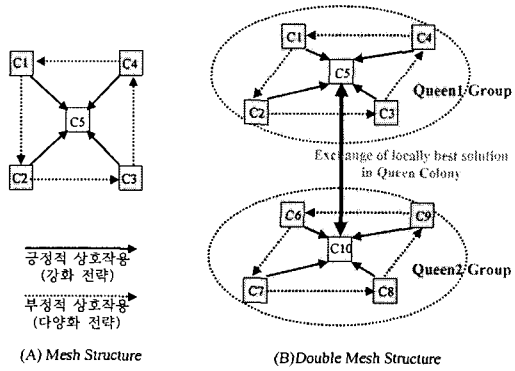


그림 1. 엘리트 전략에 의한 상호작용

다중 집단 상호작용 개미 모델에서 모든 집단은 서로 독립적으로 탐색을 수행한다. 일개미 집단간의 동작을 살펴보면, 집단l에 있는 에이전트(k)가 노드(r)에서 노드(u)로의 이동하기 위해 식(11)의 상태전이 규칙에 의해 수행된다.

$$s=\begin{cases}\arg\max_{u\in J_k^l(r)}\left\{\left[AQ^l(r,u)\right]^{\delta(l)}\cdot\left[HE^l(r,u)\right]^{\beta(l)}\right\} & ,if\ q\le q_0\\ S & ,otherwise\end{cases} \tag{11}$$

$AQ^l(r,u)$는 집단l에 있는 간선$E(r,u)$에 관계된 양의 값(positive value)이다. $AQ^l(r,u)$는 Q-학습의 Q-값와 상응하는 값으로 노드(r)에서 노드(u)로 이동하는 것이 얼마나 유용한지를 나타내는 페로몬 양으로 이동함에 따라 그 값이 갱신된다.

$HE^l(r,u)$는 집단l에 있는 에이전트가 노드(r)에서 노드(u)를 선택할 때 우수성을 평가하는 휴리스틱 값이다 (TSP문제에서는 노드(r)에서 노드(u)사이 거리의 역수). $J_k^l(r)$은 현재 집단l에서 노드(r)에 있는 에이전트(k)가 방문할 수 있는 남아있는 노드들의 집합이다.

파라미터$\delta(l)$과 $\beta(l)$은 AQ-값과 휴리스틱 값의 상대적 중요도를 나타낸다. q는 [0,1]사이에 분포된 무작위 파라미터이고, q_0 는 [0,1]사이의 값을 가지는 인자, S는 집단l에서 에이전트가 노드(r)에서 노드(s)를 선택할 때 식(12)의 확률분포에 따라서 선택된 임의 파라미터이다.

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

$$where\ \ [AQ^l(r,s)]^{\delta(l)}=\begin{cases}\sum_{n=1}^{M}[AQ^n(r,s)]^{\delta(l,n)} & ,Positive\\ \sum_{n=l-1}^{l}[AQ^n(r,s)]^{\delta(l,n)} & ,Negative\end{cases} \tag{12}$$

여기서 M은 전체 집단 수이고, $\delta(l,n)$은 집단l이 집단n으로부터 영향을 받는 상호작용의 정도를 나타내는 것으로, 일개미 집단간의 상호작용은 다양화 전략에 의한 부정적 상호 작용만 수행한다. 부정적 상호작용인 경우는 해당 간선에 대해 유방향 연결망 구조로 연결된 바로 이전 집단에서 방문한 에이전트의 방문 빈도수의 음의 역수 값(-1/해당 간선에 대해 이전 집단에서 에이전트 방문 빈도수)과 현재 집단에서 현재까지 해당 간선을 방문한 에이전트의 빈도수의 음의 역수 값(-1/현재 집단에서 해당 간선에 대해 현재까지 에이전트 방문 빈도수)의 합의 값만큼 영향을 받는다. 이 다양화 전략에 의한 부정적 상호작용을 통해 일개미 집단의 각 에이전트들은 새로운 탐색 영역으로의 다양한 탐색을 수행하게 된다.

그리고 일개미 집단과 중심 집단(여왕개미 집단)사이에는 강화 전략에 의한 긍정적 상호작용을 수행한다. 긍정적 상호작용인 경우, 중심 집단은 모든 일개미 집단으로부터 해당 간선에 방문한 에이전트의 방문 빈도수의 양의 역수 값으로 영향을 받는다. 이것은 일개미 집단에서의 모든 탐색 결과를 바탕으로 중심 집단의 에이전트들은 좋은 간선만 선택하게 되고 해당 간선을 강화하게 된다. 이것은 상호작용의 정도를 고정 값으로 할당[8]하는 것과 다르며, 우리는 방문 빈도수를 이용해 동적으로 할당한다. 따라서 $AQ^l(r,s)$는 해당 간선에 대해 각 집단간 독립적 상호작용에 의한 페로몬 영향 정도를 나타내는 변수로 표현할 수 있으며, 집단l에 속한 에이전트k는 모든 집단간 상호작용 정도에 따라, 즉 상태전이 확률에 따라 다음 노드를 선택하고 선택된 간선에 대한 AQ-값은 식(14)의 Ant-Q 기반의 다중 집단 상호작용 개미 모델에 TD-오류를 적용한 방법으로 갱신된다.

앞서 설명한 TD 학습을 이용한 Ant-Q 개미 모델을 다중 집단 상호작용 개미 모델에 적용하면 식(7)과 같다.

현재 상태의 집단l의 노드(r)에 있는 에이전트(k)에 의해 선택된 노드(s)에 대한 Q-값($AQ^l(r,s)$)과 현재 상태의 노드(r,s)에 의해 선택된 다음 상태의 노드(s,z) 중에서 최대 Q-값($MaxAQ^l(s,z)$)을 갖는 노드(s,z) 쌍과의 Q-함수 값을 갱신하기 위해 TD 학습을 이용한다. TD 학습은 식(13)과 같이 계산된다.

$$TD\ error = \Delta AQ^l(r,s) + \gamma \cdot [\underset{z \in J_k^l(s)}{Max}\ AQ^l(s,z) - AQ^l(r,s)] \quad (13)$$

결국, Ant-Q 기반의 다중 집단 상호작용 개미 모델에 TD 학습을 적용한 다중 집단 상호작용 개미 강화학습 모델은 TD 학습을 이용하여 최적의 값-함수를 구하기 위해 상태전이 이후 현재 상태의 노드(r,s) 쌍에 대한 Q-함수 값을 식(14)와 같이 생신한다.

$$\begin{aligned} AQ^l(r,s) \leftarrow & (1-\alpha)\cdot AQ^l(r,s) \\ & + \alpha \cdot (\Delta AQ^l(r,s) + \gamma \cdot [\underset{z \in J_k^l(s)}{MaxAQ^l(s,z)} - AQ^l(r,s)]) \end{aligned}$$

$$\text{where}\quad \Delta AQ^l(r,s) = 0 \qquad ,if\ Local\ updating$$

$$\underset{z \in J_k^l(s)}{MaxAQ^l(s,z)} - AQ^l(r,s) = 0 \quad ,if\ Global\ updating \quad (14)$$

$\alpha(0<\alpha<1)$는 페로몬 지연 파라미터로 학습율, γ 는 할인율이다. Local updating 이란 에이전트들이 상태과정을 수행하는 동안 간선에 페로몬을 갱신하는 과정이며, Global updating 이란 에이전트들이 경로 사이클을 완성 후 간선에 페로몬을 갱신하는 과정이다. $MaxAQ^l(s,z)$는 다음 상태에 대한 평가로 외부 환경으로부터 받는 강화값을 최대화하는 것으로 전역 강화일 때는 0이다.

$\Delta AQ^l(r,s)$는 강화값으로 지역 강화일 때는 항상 0이다. 전역 강화는 에이전트들이 모든 경로를 완성 후에 다음의 식(15)에 의해 갱신된다.

$$\Delta AQ^l(r,s) = \begin{cases} \dfrac{W}{L_{k_{lb}^l}} & ,if\ (r,s) \in tour\ done\ by\ the\ agent\ k_{gb}^l \\ 0 & ,otherwise \end{cases} \quad (15)$$

W는 상수 값으로 여러 실험을 통해 $W=10$으로 고정한다. $L_{k_{lb}^l}$는 집단l의 현재 경로 사이클의 최적 경로 길이이다. 또한 Queen1그룹의 중심 집단(C5)과 Queen2 그룹의 중심 집단(C10)간의 상호작용은 두 그룹의 지역 최적해를 서로 비교해, 우수한 해에 대해 전역 갱신하는 전략을 채택한다.

Ⅳ. 다중 집단 개미 강화학습 모델 성능 평가

TD 학습을 적용한 Ant-Q 기반의 다중 집단 상호작용 개미 강화학습 모델의 성능을 평가하기 위해, 실험은 네 가지 방향으로 진행한다. 첫째, 학습율(α)에 따른 탐색 결과의 변화를 관측함으로써 학습율에 따른 영향을 평가한다. 두 번째는 할인율(γ)에 따른 영향을 평가한다. 세 번째는 q_0에 따른 영향을 평가한다. 그리고 네 번째는 기존 개미 모델과의 탐색 결과 비교를 통한 성능을 측정하여 본다. 제안된 개미 모델을 실험하기 위해서 도시들의 위치는 TSP 예제로 널리 알려진 TSPLIB[10]에서 추출하여 실험을 하였다. 실험을 위한 개미 모델의 기본 환경 변수는 다음과 같이 결정하였다.

$\delta(l,n)=\pm(1/$방문 빈도수$)$, $\beta(l)=2$, $\alpha=0.1$, $\gamma=0.3$, $q_0=0.9$, $\tau_0=(n*L_{nn})^{-1}$, $W=10$, $m=10$. 각 집단에서 에이전트들의 초기 위치 결정은 각 노드에 1개씩 무작위로 배정하였으며, 종료 조건은 고정된 수행 횟수 또는 여러 실험에 의해 최적해로 알려진 값을 찾았을 경우 종료하였다.

[그림 2]는 Mesh 구조에서 사이클 횟수를 2000으로 했을 때, 다양한 노드 집합에서의 성능을 평가하고 있다. 학습율이 0.1일 때 가장 좋은 성능을 보이며, 학습율이 높아질수록 성능이 점차 감소됨을 실험에서 밝히고 있다. 여기서 Eil51.TSP 문제는 산출된 수치 그대로이며, St70.TSP 문제는 동일 그래프상에 표현하기 위해

산출된 결과의 62% 수치 값으로 표현하였다.

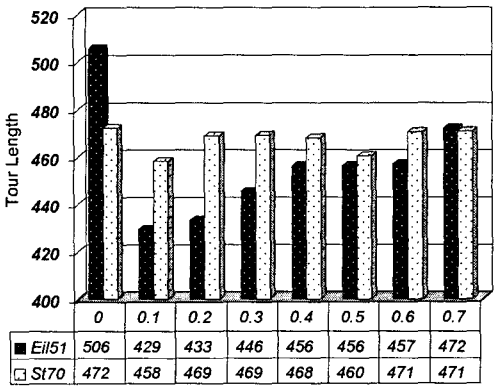


그림 2. 학습율(α)에 따른 성능 평가

[표 1]은 제안된 개미 모델을 이용해 Eil51.TSP 문제에서의 할인율(γ) 변화에 따라 사이클 횟수, 시간 그리고 경로 길이를 평가하고 있다. 그 결과를 살펴보면, 대부분의 문제에서 할인율이 0.1~0.3 구간에서 좋은 탐색 결과를 보이고 있다. 특히 실험에 사용된 문제에서 할인율(γ)을 0.3으로 설정해 탐색하는 것이 최종적인 탐색 결과를 우수하게 만들 수 있다는 것을 실험에서 밝히고 있다.

[표 2]는 다양한 문제 영역에서의 할인율에 따른 탐색 결과를 수치화한 것으로 할인율이 0.3일 때 가장 좋은 탐색 성능을 보이고 있다.

표 1. 할인율(γ)에 따른 평가

할인율 (γ)	사이클 (Cycle)	시간 (Second)	Proposed Ant Model	Ant-Q
0	872	540.64	432.57	436.28
0.1	1036	642.32	429.53	430.35
0.2	1178	730.36	430.88	430.40
0.3	**74**	**45.88**	**429.48**	**430.40**
0.4	318	197.16	429.74	431.97
0.5	137	84.94	430.35	431.61
0.6	639	396.18	431.11	432.33
0.7	1613	1000.06	431.98	434.59
0.8	195	120.9	429.53	430.35
0.9	1756	1088.72	433.94	435.75
1	248	153.76	431.98	433.6

표 2. 다양한 문제 영역에서의 할인율(γ)에 따른 평가

할인율 (γ)	Node Set			
	Eil51	St70	Rat99	KroA100
0	432.57	689.23	1256.79	22301.78
0.1	429.53	677.44	1221.01	22013.64
0.2	430.88	677.44	1228.63	21815.59
0.3	**429.48**	**677.44**	**1219.24**	**21422.82**
0.4	429.74	683.70	1254.97	21808.84
0.5	430.35	687.22	1261.25	22354.65
0.6	431.11	688.02	1255.58	22587.87
0.7	431.98	687.35	1268.21	22519.18
0.8	429.53	686.32	1266.32	22646.79
0.9	433.94	689.69	1265.24	22702.70
1	431.98	690.69	1267.88	22808.65

[그림 3]은 Eil51.TSP 문제를 이용해 Ant-Q개미 모델과 제안된 개미 모델에서 q_0에 따른 성능 평가 결과를 보여주고 있다. 결과를 살펴보면, q_0가 0.7~0.9사이에서 좋은 결과를 보여주고 있으며, 그 중에서 q_0가 0.9일 경우 가장 좋은 결과를 산출함을 볼 수 있다.

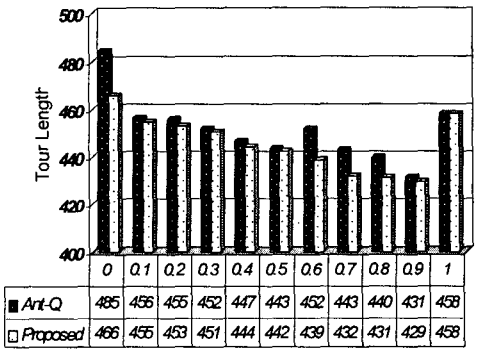


그림 3. q_0에 따른 성능 평가

[표 3]은 Double Mesh 구조에서 ACS 개미 모델, Ant-Q 개미 모델, 다중 집단 상호작용 개미 강화학습 모델을 동일 환경에서 수행한 실험 결과이며, 사이클 횟수는 20000회의 결과이다. 그 결과를 살펴보면, 각 ACS, Ant-Q 그리고 제안된 개미 모델에 의해 산출된 최적 경로 길이와 평균 경로 길이를 보여주는 것으로 제안된 개미 모델의 성능이 우수하다는 것을 보여주고

있다.

이것은 문제 영역이 커질수록 다양한 상태전이를 통한 다양화 전략과 강화 전략, 그리고 TD 학습을 이용한 방법이 효과적임을 보여주고 있다. 그러나 Ant-Q 개미 모델이 계산량이 많아 시간이 많이 소요되는 이유로 제안된 개미 모델도 그에 따른 계산량이 많아지는 단점이 있음에도 불구하고 문제영역이 큰 문제에 대해 제안된 개미 모델이 효과적으로 적용될 수 있음을 실험을 통해 알 수 있었다.

표 3. 다중 집단 상호작용 개미 강화학습 모델 평가

Node Set	ACS[9]		Ant-Q[4]		Proposed Model	
	Average Length	Best Length	Average Length	Best Length	Average Length	Best Length
KroA150	28908.8	27824	28761.42	27231	26576.04	26524
Rat195	2571.63	2461	2514.08	2397	2490.25	2338
Gil262	2636.75	2526	2592.35	2493	2481.68	2389
A280	2892.58	2768	2840.52	2758	2690.61	2585
Pr299	53497.8	51395	52714.41	50278	49143.87	48312
Lin318	46244.4	44837	45318.53	43832	44571.98	43041

V. 결 론

본 논문에서는 TD 학습을 적용한 Ant-Q 기반의 다중 집단 상호작용 개미 강화학습 모델을 제안하였다.

본 논문에서 제안된 개미 모델 학습 방법은 기존의 Ant-Q 개미 모델 학습 성능을 개선하기 위해 새롭게 제안된 방법이다. 이 방법은 Ant-Q 개미 모델에 TD 학습을 통한 강화학습과 엘리트 전략에 의한 다중 집단 상호작용 개미 모델을 적용한 학습 방법이다.

TD 학습을 이용한 제안된 개미 모델은 매 학습 단계에서 현재 상태의 출력에 대한 예측과 다음 상태의 출력에 대한 예측과의 차이를 이용하여, 현재 상태에서 현재 상태의 출력에 대한 예측과 다음 상태의 출력에 대한 예측과 가깝게 하기 위해 갱신하였다. 그리고 엘리트 전략에 의한 집단간 상호작용을 통해 각 에이전트들이 경로 사이클을 이루는 동안 각 간선에 방문한 방문 빈도수 기반의 다양화 전략을 상태전이 규칙에 적용해 에이전트들이 탐색영역을 더욱 다양하게 검색 가능하게 하였다. 이로 인해 에이전트는 선호하지 않는 새로운 탐색 영역으로의 탐색 공간 확장을 통해 국부 최적으로부터 벗어날 수 있었고, 또한 최적해에 빠르게 수렴할 수 있었다.

실험은 학습율과 할인율 그리고 q_0의 변화에 따른 탐색 결과의 변화를 측정하였고, 제안된 TD 학습을 이용한 다중 집단 개미 강화학습 모델의 성능을 실험하였다.

향후 연구과제는 제안된 개미 모델에서 현재 상태에서 선택한 노드에 대해 얼마나 적합한가를 의미하는 척도인 적합도(Eligibility factor)를 이용한 강화학습 방법에 대한 연구도 필요하다.

참 고 문 헌

[1]. 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.

[2]. 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.

[3]. T. Stutzle and H. Hoos, "The ant system and local search for the traveling salesman problem," Proceedings of ICEC '97-1997 IEEE 4th International Conference of Evolutionary.

[4] 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.

[5] 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.

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

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

[8] H. Kawamura, M. Yamamoto, K. Suzuki, and A. Ohuchi, "Multiple Ant Colonies Algorithm Based on Colony Level Interactions," IEICE Transactions, Vol.E83-A, No.2, pp.371-379, 2000.

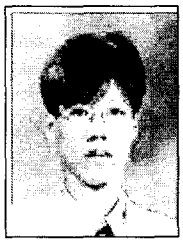
[9] L. M. Gambardella and M. Dorigo, "Ant Colony System: A Cooperative Learning approach to the Traveling Salesman Problem," IEEE Transactions on Evolutionary Computation, Vol.1, No.1, 1997.

[10] http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/

[11] S. G. Lee, "Multiagent Reinforcement Learning Algorithm Using Temporal Difference Error," Springer-Verlag, Berlin, LNCS, Vol.3496, pp.627-633, 2005.

저 자 소 개

이 승 관(Seung-Gwan Lee) 정회원



- 1997년 2월 : 경희대학교 전자계산공학과(공학사)
- 1999년 2월 : 경희대학교 전자계산공학과(공학석사)
- 2004년 2월 : 경희대학교 전자계산공학과(공학박사)
- 2004년 3월~현재 : 가톨릭대학교 컴퓨터정보공학부 강의 전임 교수

<관심분야> : 인공지능, 로봇에이전트, 최적화, 데이터마이닝, 유비쿼터스 컴퓨팅