# 강화학습의 Q-learning을 위한 함수근사 방법
## (A Function Approximation Method for Q-learning of Reinforcement Learning)

이 영 아 † 정 태 충 ††
(YoungAh Lee) (TaeChoong Chung)

**요 약** 강화학습(reinforcement learning)은 온라인으로 환경(environment)과 상호작용 하는 과정을 통하여 목표를 이루기 위한 전략을 학습한다. 강화학습의 기본적인 알고리즘인 Q-learning의 학습 속도를 가속하기 위해서, 거대한 상태공간 문제(curse of dimensionality)를 해결할 수 있고 강화학습의 특성에 적합한 함수 근사 방법이 필요하다. 본 논문에서는 이러한 문제점들을 개선하기 위해서, 온라인 퍼지 클러스터링(online fuzzy clustering)을 기반으로 한 Fuzzy Q-Map을 제안한다. Fuzzy Q-Map은 온라인 학습이 가능하고 환경의 불확실성을 표현할 수 있는 강화학습에 적합한 함수근사방법이다. Fuzzy Q-Map을 마운틴 카 문제에 적용하여 보았고, 학습 초기에 학습 속도가 가속됨을 보였다.

**키워드** : 강화학습, Q-learning, 함수근사, 온라인 퍼지 클러스터링

***Abstract*** Reinforcement learning learns policies for accomplishing a task's goal by experience through interaction between agent and environment. Q-learning, basis algorithm of reinforcement learning, has the problem of curse of dimensionality and slow learning speed in the incipient stage of learning. In order to solve the problems of Q-learning, new function approximation methods suitable for reinforcement learning should be studied. In this paper, to improve these problems, we suggest Fuzzy Q-Map algorithm that is based on online fuzzy clustering. Fuzzy Q-Map is a function approximation method suitable to reinforcement learning that can do on-line learning and express uncertainty of environment. We made an experiment on the mountain car problem with Fuzzy Q-Map, and its results show that learning speed is accelerated in the incipient stage of learning.

**Key words** : reinforcement learning, Q-learning, function approximation, online fuzzy clustering

## 1. 서 론

학습 방법은 지도 학습(supervised learning), 비지도 학습(unsupervised learning), 강화학습(reinforcement learning)[1-3]으로 나눌 수 있다. 지도 학습은 모든 입력에 대하여 정확한 답을 가지고 학습하는 방법이다. 비지도 학습은 입력에 대한 정확한 답이 없이 입력 데이타에 내재된 구조나 관계를 파악하고, 이 관계를 이용해서 입력 패턴들을 분류한다. 강화학습은 환경(environment)과의 상호 작용하는 과정에서 선택한 행동의 좋고 나쁨에 따라 주어지는 보답(reward)을 이용하여 보다 좋은 결과를 얻을 수 있는 행동을 학습한다. 에이전트는 환경이 제시하는 훈련 데이타를 통해서 원인과 결과(cause and effect), 행동의 결과, 목표를 성취할 수 있는 행동에 대한 정보를 학습하게 된다.

강화학습의 목적은 불확실한 환경에서 에이전트(agent)가 환경과의 상호 작용하는 과정에서 시행착오(trial-and-error)를 통하여 목적(goal)에 도달할 수 있는 가치 함수(value function)를 학습하는 것이다. 가치 함수는 상태의 가치(value) 또는 장 기간의 가치(long-term utility)를 평가하는 함수로서, 에이전트가 다음 행동을 결정할 때 사용될 수 있다.

강화학습의 기본 알고리즘인 Q-Learning[1-4]은 상태-행동의 가치 함수(state-action value function)인 Q-함수를 계산하는데, 이 가치 함수는 한 단계를 미리 예측하고 그 결과를 최적의 전략의 계산에 함께 이용한다. 기본 알고리즘에서 Q-함수는 테이블의 형태에 저장되고 상태와 행동에 의해서 참조된다.

Q-learning은 몇 가지 단점을 가지고 있다.

첫째, 상태와 행동이 연속적이거나 복잡한 문제인 경

† 학생회원 : 경희대학교 컴퓨터공학과
leeyaa@iislab.kyunghee.ac.kr
†† 종신회원 : 경희대학교 컴퓨터공학과 교수
tcchung@nms.kyunghee.ac.kr
논문접수 : 2004년 6월 10일
심사완료 : 2004년 9월 9일

우에는 상태공간 또는 행동공간의 크기가 거대해서 메모리에 모든 상태-행동 쌍을 저장할 수 없고, 학습 시간이 길어지며 모든 상태-행동 쌍을 경험할 수 없다. 이러한 문제를 "curse of dimensionality" 문제라고 한다.

둘째, 실세계에서 동작하는 에이전트는 시뮬레이션과는 달리 실제적으로는 많은 양의 훈련 데이타를 얻을 수 없다.

셋째, Q-learning의 보상함수(reward function)는 오직 목표 상태(goal state)에서만 보상을 주고 나머지 상태들에는 같은 벌금값(penalty)을 부여하는 간단한 형태가 선호된다. 그러한 보상함수를 이용한다면, 에이전트는 목표 상태에 도달할 때까지 임의적으로 행동을 선택하여야 한다. 온라인 학습에서 중요한 것은 적은 양의 훈련을 거쳐서 적절한 예측을 해야 하는 것인데, 목표에 도달할 때까지 임의적으로 행동을 선택해야 하므로 학습 속도는 느려지게 된다.

함수 근사 방법(function approximators)[1-4]은 유사한 상태 또는 행동들을 일반화함으로써 Q-함수를 압축할 수 있다. 모든 상태에서 최적의 행동을 알기 위해서는 각 상태에서 가능한 모든 행동을 한번 이상은 시도해야 하지만 함수 근사 방법은 방문하지 않은 상태-행동 쌍을 일반화함으로써 문제를 해결한다.

강화학습의 함수 근사 방법이 갖추어야 할 몇 가지 요구 사항은 다음과 같다.

첫째, 강화학습은 온라인 학습 방법이므로 한번에 하나의 데이타를 학습하는 것이 가능하여야 하고, 적은 수의 훈련 데이타만을 바탕으로 일찍부터 적절한 예측을 하여야 한다.

둘째, 강화학습은 상태 공간을 탐험하는 경로를 따라서 훈련 데이타가 생성되므로, 훈련 데이타의 분포가 시간에 따라서 변한다. 따라서 최근에 방문하지 않은 공간의 학습 내용은 파괴될 수 있다. 그러한 파괴적 간섭(destructive interference)현상이 없어야 한다.

셋째, 강화학습의 불확실성의 원인은 동적인 환경과 환경에 대한 모델이 없기 때문이다. 강화학습의 불확실성을 표현할 수 있어야 한다.

만일 사전지식이 충분하게 주어지지 않고 거대한 상태공간을 갖는 강화학습 문제를 해결해야 한다면, 환경의 정황을 표현하는 상태들을 유사한 정도에 따라서 분류할 수 있어야 하고, 각 군집에서 가능한 행동들을 학습할 수 있어야 하며, 분류된 각 군집은 환경의 변화에 따라서 적응할 수 있어야 한다. 그리고 적은 수의 훈련 데이타를 경험한 후에 적절한 응답을 할 수 있을 만큼 학습속도가 가속화되어야 한다. 이러한 조건을 만족시킬 수 있는 함수 근사 방법은 온라인 퍼지 클러스터링[5-13]을 기반으로 하는 것이 적합하다고 본다.

본 논문에서는 온라인 퍼지 클러스터링을 바탕으로 사전 지식의 추출이 어렵다는 가정 하에, 연속적인 상태공간이 갖는 문제를 개선하고 학습 속도를 가속화하여 빠른 시간에 타당한 예측이 가능한 함수 근사 방법인 Fuzzy Q-Map을 제안한다.

## 2. 관련 연구

### 2.1 Q-learning

Q-learning[1-4]은 Watkins가 처음 제안한 강화학습의 대표적인 알고리즘으로서, 상태-행동의 가치 함수인 Q-함수를 학습한다. 함수 $Q(s,a)$는 장기간의 관점에서 봤을 때, 상태 $s$에서 행동 $a$을 선택하는 것이 어느 정도 유익한 가를 표시한다. Q-learning은 온 라인으로 가치함수와 전략을 동시에 학습할 수 있다. Q-value는 임의로 초기화 되고, 다음 식에 따라 갱신된다. 반복적으로 $Q(s,a)$를 근사한다면, 최적의 $Q^*(s,a)$으로 수렴한다.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)) \quad (1)$$

Q-learning은 모든 상태-행동 쌍을 하나의 룩업 테이블에 저장하고 반복해서 경험해야 하므로, 상태공간의 크기에 따라서 메모리의 필요량이 증가하고 학습 시간이 증가한다.

강화학습의 보상함수는 일반적으로 오직 목표 상태로 전이한 경우에만 보상값을 부여하고 목표가 아닌 나머지 상태가 다음 상태가 된다면 같은 값의 벌금값을 부여하는 간단한 형태이다. 이러한 보상함수를 이용해서 학습한다면, 행동 선택에 지표가 없으므로 목표상태에 도달하기 전까지 임의로 행동을 선택하여야 한다.

### 2.2 강화학습의 학습 속도 개선을 위한 연구

강화학습과 같은 온라인 학습 방법은 입력으로 들어온 훈련 데이타에 대하여 적절한 대응을 하여야 한다. 만일 간단한 형태의 보상함수를 사용하고 사전지식이 주어지지 않는다면, 강화학습 에이전트는 학습 초기에는 동일하게 초기화된 Q값과 적은 경험을 바탕으로 행동 선택을 해야 한다. 강화학습은 현재 시점까지의 Q-함수를 이용해서 행동을 선택하고, 그 결과를 이용해서 Q-함수를 갱신하므로 초기의 적절한 선택이 학습 속도를 가속화할 수 있다. 강화학습의 학습 속도 개선을 위한 연구는 다양하게 이루어지고 있다.

#### 2.2.1 Fuzzy Q-Learning

Fuzzy Q-Learning(FQL)[14-17]은 Glorennce와 Jouffe가 Watkins의 Q-Learning에 퍼지 추론 시스템(Fuzzy Inference System, FIS)을 적용한 학습 방법이다. 퍼지 추론 시스템은 연속적인 값을 갖고 불확실성을 포함한 복잡한 모델을 학습하는데 적합하고 문제(task)

에 대한 사전 지식을 쉽게 규칙에 표현할 수 있다. 연속값은 유한개의 구간으로 나누어 처리될 수 있는데, 이때 적당한 구간으로 나누는 것은 쉬운 일이 아니다. 또한 입력 상태가 어떠한 구간에 속하더라도 구간 자체의 경계가 명확하지 않다. 이러한 시스템에서, 입력 상태를 어떠한 구간과 1:1 매칭한다면 에러가 발생하고, 훈련하는 과정에서 에러의 크기는 점차적으로 증가한다. 이러한 문제를 퍼지 이론을 바탕으로 모델링 한다면 융통성 있는 학습이 이루어지므로 에러의 크기를 줄일 수 있는 장점이 있다. FQL은 행동과 Q 값을 퍼지 규칙으로부터 추론 한다. FQL은 학습 속도의 개선이라는 성과를 거두었지만, 퍼지 레이블과 소속도 함수를 결정하기 위한 사전 정보와 사전처리가 필요하다. 그리고 퍼지 규칙의 조건부(condition part)는 학습을 시작하기 전에 결정되므로 강화학습 학습 과정에서는 고정되는 문제점이 있다.

2.2.2 CMAC(Cerebellar Model Articulation Controller)

Sutton이 제안한 CMAC(Cerebellar Model Articulation Controller)[1,18-20]는 퍼지이론을 사용하지는 않았지만 상태공간을 일부 겹치는 여러 개의 타일링(tiling)을 운영하여 행동선택에 융통성을 주고 있다. CMAC는 입력 상태에 대하여 각 타일링에서 활성화된 타일들(tiles)의 평균을 학습에 이용한다. 질의 $x_q$가 들어오면, 질의에 대한 Q값은 다음 식과 같이 활성화된 타일들의 집합 $F(x_q, a_q)$에 포함된 타일들의 가중치 합이다.

$$Q'(x,a)=\sum_{f_{ij}\in F(x,a)} w_{ij} \tag{2}$$

CMAC의 모든 타일의 가중치 $w_{ij}$는 다음 식 3과 같이 갱신된다. 적합도 $e_f$는 다음 식 (4)와 같이 갱신된다.

$$\triangle w_{ij}=\alpha(r_t+\gamma Q'_{t+1}-Q'_t)e_f \qquad \forall f_{ij}\in CMAC \tag{3}$$

$$e_{ij}\leftarrow\begin{cases}\dfrac{1}{|F(x_q,a_q)|} & \text{if } f\in F(x_q,a_q)\\ \lambda\gamma e_{ij} & otherwise\end{cases} \tag{4}$$

$TD(\lambda)$와 적합도의 갱신은 CMAC의 타일의 수만큼 이루어지므로 타일과 타일링의 수가 많을수록 일반화된 좋은 결과를 얻을 수 있지만 기억장소도 비례해서 증가한다.

2.2.3 LWR(Locally Weighted Regression)

훈련 예제가 주어진다면 이를 효율적으로 학습에 이용하는 연구가 필요하다. 또한 훈련 예제가 거의 없는 경우 학습 성능을 향상시키는 연구도 필요하다. 사전에 주어진 훈련 예제는 목표에 도달할 수 있는 행동을 선택하는데 이용될 수 있으므로 강화학습의 학습 속도의 개선을 이룰 수 있다. 훈련 예제는 '상태-행동'의 형태로 특정한 상황을 표현하는 '상태'와 그 상태에서 최적이거나 최적에 가까운 '행동'으로 이루어진다. 훈련 예제는 도메인을 잘 아는 전문가에 의하여 수집될 수 있다. 하지만 사전지식과 사용자의 지도를 이용하는 시스템들은 학습 속도를 개선하지만 자율적인 학습이라고 볼 수 없고, 사전지식은 항상 쉽게 수집될 수 있는 것이 아니다.

Smart는 국소 최소 자승법 (Locally Weighted Regression: LWR)을 기본으로 한 함수 근사 방법인 HEDGER와 JAQL[4,21,22]을 소개하였고, Aljibury는 먼저 적은 수의 상태들을 Q-learning 알고리즘으로 학습하고 다음 단계에서 LWR을 적용하여 연속적인 상태공간의 가치 함수로 일반화 시켰다. LWR은 커널 함수(kernel function)를 이용하여 질의(query point)와 거리가 가까운 훈련 예제들이 예측에 보다 큰 영향을 주도록 하고 지역적인 함수를 구하는 방법이다. LWR에서 사용하는 커널함수는 다음과 같이 가우시안 함수이다. $h$는 함수의 완만함을 조절하는 파라미터이고, 입력 데이타와 훈련 예제사이의 유클리드 거리에 따라서 가중치 $k_i$가 결정된다.

$$k_i=e^{-(\frac{d_i}{h})^2} \tag{5}$$

### 2.3 퍼지 클러스터링의 소속도

위의 알고리즘들은 입력 상태들을 유사한 정도에 따라서 분류하는데, 유사한 정도를 구하는 소속도 함수는 퍼지 클러스터링 에서 다음과 같이 상대적 소속도(relative membership degree) $R_{i,k}$와 절대적 소속도(absolute membership degree) $A_{i,k}$로 나눈다.

두 소속도의 관계는 다음과 같은 식 (6)으로 표현할 수 있다.

$$R_{i,k}=\frac{A_{i,k}}{\sum_{h=1}^{c} A_{h,k}},\ i=1,\ldots,c,\ k=1,\ldots,n \tag{6}$$

$k$는 데이타에 대한 인덱스이고, $i$는 군집에 대한 인덱스이다. 상대적인 값 $R_{i,k}$는 다음의 세 가지 조건을 만족하여야 한다.

$$(i)\ R_{i,k}\in(0,1],\ i=1,\ldots,c,\ k=1,\ldots,n \tag{7}$$

$$(ii)\ \sum_{i=1}^{c} R_{i,k}=1,\ k=1,\ldots,n \tag{8}$$

$$(iii)\ 0<\sum_{k=1}^{n} R_{i,k}<n,\ i=1,\ldots,c. \tag{9}$$

제한 조건 $(ii)$에 의하여, $R_{i,k}$는 다른 모든 클래스의 절대적 소속도에 따른 상대적인 값이 된다. 둘째 제한 조건에 의하여, 노이즈나 이상치도 높은 소속도를 가질 수 있고 전체적인 모델의 특성 평가에 영향을 줄 수 있다.

절대적 소속도 $A_{i,k}$는 주어진 데이타가 하나의 군집에 속하는 유사 값을 배타적으로 표현한다. $A_{i,k}$는 $R_{i,k}$의 둘째 제한 조건을 다음과 같이 완화한다.

$$(vi)\ \max_i A_{i,k} > 0,\ k = 1, \ldots, n \quad (10)$$

새로운 조건에 의하여, 잡음 데이타에 대한 절대적 소속도의 합이 1이 될 필요가 없어진다. 그러나 절대적 소속도의 단점은 각 군집에서 독립적으로 목적 함수의 최소화를 시도하므로 지역적 극소점의 수가 증가하게 된다는 것이다.

Fuzzy c-means(FCM)와 Fuzzy Learning Vector Quantization(FLVQ)은 상대적 퍼지 소속도를 이용한다. 이는 입력 데이타와 군집의 관계에 대하여 승자와 승자가 아닌 다른 군집들의 모든 정보를 바탕으로 하는 것이 지역적인 정보를 바탕으로 하는 것보다 전체적인 입력공간을 보다 잘 표현할 수 있기 때문이다.

본 논문에서 제안한 Fuzzy Q-Map은 상대적인 퍼지 소속도를 바탕으로 입력 공간을 분석하고, 행동 값을 제한하였으며 전역적인 전략을 학습하고자 한다.

## 3. Fuzzy Q-Map

### 3.1 강화학습에 적합한 함수 근사 방법

강화학습의 거대한 상태 공간 문제를 해결하기 위해서, 상태 공간을 탐험하는 과정에서 만나지 못한 상태들은 이미 경험한 상태들로부터 일반화하여야 한다. 이러한 일반화를 함수 근사(function approximation)라고 한다. 함수 근사는 지도 학습에서 이미 널리 연구되었는데, 지도학습은 주어진 훈련 예제 집합을 반복적으로 학습하는 방법으로 강화학습과 차이가 있다. 강화학습은 에이전트가 선택한 행동에 따라서 탐험 경로가 달라지고, 탐험 과정에서 만나는 상태들을 경험하게 된다. 그리고 환경의 변화에 대하여 짧은 시간에 적절한 반응을 해야 한다. 이러한 차이점에 의하여, 기존의 함수 근사 방법을 그대로 강화학습에 적용할 수 없고 강화학습의 특징을 고려한 함수 근사 방법에 대한 연구가 필요하다. 본 논문에서는 강화학습에 적합한 함수 근사 방법인 온라인 퍼지 클러스터링과 Q-learning을 결합한 Fuzzy Q-Map을 제안한다. 온라인 퍼지 클러스터링이 강화학습에 적합한 이유는 다음과 같다.

첫째, 강화학습은 교사가 없다는 점에서, 비지도 학습 방법인 클러스터링과 유사하다. 클러스터링은 유사한 상태들은 그룹화하고, 계속해서 들어오는 데이타에 적응할 수 있다. 둘째, 퍼지 클러스터링은 학습 속도의 가속화와 연속적인 상태 공간을 표현할 수 있다. 강화학습은 환경에 대한 모델이 없으므로 구간 자체의 경계가 명확하지 않다. 만일 입력 상태를 어떠한 구간과 1 : 1 매칭한다면 에러가 발생하고, 훈련 하는 과정에서 에러의 크기는 점차적으로 크게 증가한다. 퍼지 클러스터링의 소속도 함수는 이러한 불확실성을 표현한다. 훈련 데이타와 유사한 군집들은 소속도 만큼 전략을 제시하고 전략을 갱신하므로 에러의 크기는 감소하고 학습 속도는 가속화된다. 셋째, 온라인 알고리즘은 실시간의 응답을 하고 가장 최근에 입력된 데이타만이 순차적인 갱신에 영향을 주는 방법이다. 데이타는 사용되고 나면 시스템에서 제거가 되므로 온 라인 클러스터링은 거대한 훈련 데이타 집합 전체를 메모리에 기억할 필요가 없고, 최근의 데이타에 대한 학습만 하므로 계산량도 감소한다. 그리고 개념의 변화에 따른 데이타의 변화에도 적응할 수 있다. 넷째, 클러스터링은 지역적인 학습 방법이다. 전략은 복잡한 전역적인 함수가 아닌 간단한 지역적 전략들의 조합으로 표현된다. 그러므로 신경망과 같은 파괴적인 간섭현상을 피할 수 있다.

### 3.2 Fuzzy Q-Map의 구조

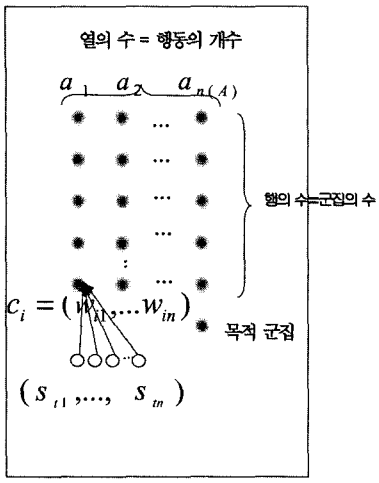
Fuzzy Q-Map은 각 퍼지 클러스터의 중심과 행동들의 Q 값을 기억하는 구조이다. Fuzzy Q-Map은 2차원 구조의 표 형태이다. 다차원의 입력은 거리와 퍼지 상수에 따라서 여러 군집에 소속되게 된다. 행의 수는 사용자가 지정한 군집의 개수이고, 열의 수는 상태공간에서 가능한 행동의 개수와 같다. 강화학습을 하는 동안, 에이전트는 많은 에피소드들(임의의 상태에서 출발하여 목표 상태에 도달할 때까지의 경로)을 경험하게 되는데, 한 에피소드의 끝인 목표 상태는 다음 상태로 전이되지 않고 행동들의 Q값을 학습할 필요가 없으므로 군집된 다른 상태와 같이 다룰 수 없다. 그러므로 목표 상태를 중심으로 하는 군집을 추가해서 Fuzzy Q-Map을 구성하는 노드의 수는 (군집의 수 + 행동의 수) + 1이 된다.

Fuzzy Q-Map은 이산 행동을 다룰 수 있고 또한 이산 행동을 소속도 만큼 고려함으로써 연속적인 행동도 다룰 수 있다.

각 퍼지 클러스터는 그림 1과 같이 중심과 행동, 그 행동을 수행해서 받은 보상값, 현재의 Q값을 기억하고 있다. 중심 $c_i = (w_{i1}, \ldots, w_{in})$은 상태와 같은 $n$차원의 벡터이고 군집에 배당되는 입력 상태에 적응한다.

에이전트는 예측할 때, Fuzzy Q-Map의 Q 값을 직접적으로 참조할 수 없다. Fuzzy Q-Map은 Q 값 갱신에서 소속도 만큼 적용하므로, 저장되어 있는 Q 값은 부분적인 값이다. 행동을 예측할 때, 부분적인 Q 값을 예측을 위한 정보로 이용하므로 각 군집이 제안하는 행동 또한 소속도 만큼 참조해야 한다. Fuzzy Q-Map의 행동과 Q 값을 지역 값과 전역적인 값으로 나누어 정리하면 다음과 같다.

- 지역 값(local value)
  - 최적의 행동 : 각 군집이 제안하는 행동으로 Q 값이 가장 크다.
  - Q 값 : 각 군집의 행동들에 대한 Q 값은 군집의 내부에서 행동의 가치를 평가하는 값이고 비교에




**군집의 정보**

| 중심 | 행동 | 보상값 | 지역적Q 값 |
|---|---|---|---|
| $c_i$ | $a_j$ | $r$ | $q(c_i, a_j)$ |

n : 상태의 특성 수

$(s_{t1}, \ldots, s_{tn})$: 입력 상태

$c_i = (w_{i1}, \ldots, w_{in})$: 군집 $i$의 중심

$q(c_i, a_j)$: 군집 $i$의 행동 $a_j$의 지역적인 Q 값

그림 1 Fuzzy Q-Map

사용된다.

• 전역 값(global value)

- 최적의 행동 : 입력에 대하여 Fuzzy Q-Map이 예측한 최적의 행동이다. 각 군집이 제안하는 행동을 소속도 만큼 참조한다.
- Q 값 : Fuzzy Q-Map이 상태-행동 쌍 또는 상태를 평가한 값이다.

### 3.3 Fuzzy Q-Map 알고리즘

알고리즘을 단계별로 정리하면 다음과 같다.

단계 1: (초기화)

각 군집의 중심 $c_i\,(1 \le i \le c)$을 임의적으로 초기화하고, 각 행동에 대한 보상값과 Q 값은 0으로 초기화 한다.

단계 2: (시작 상태의 선택)

입력 상태 $s_t$를 랜덤하게 선택한다. $t$는 하나의 상태를 처리하는 시점을 표시하며 지금까지 훈련에 사용된 상태의 개수를 나타낸다. 훈련 데이타 집합은 에피소드들의 집합이라고 말할 수 있는데, 즉 상태공간을 탐험하는 과정에서 만나게 되는 상태들로 이루어져 있다. 그러므로 $s_t$는 $t$ 시점에서의 한 에피소드를 구성하는 입력 상태이다.

단계 3: (소속도 $m_{it}$을 구한다.)

$s_t$가 각 군집 $i(1 \le i \le c)$에 속하는 소속도 $m_{it}$을 구한다. 소속도 $m_{it}$를 계산은 식 (11)과 같다.

$$m = \frac{1}{\sum_{j=1}^{c} \left(\frac{d}{d_{jt}}\right)^{2/(f-1)}} \tag{11}$$

$$\sum_{i=1}^{c} m = 1 \tag{12}$$

식 (11)에서 $f$는 퍼지 상수로 사용자가 지정하는 값이다. 소속도 $m_{it}$는 입력 데이타가 모든 군집에 대하여 상대적으로 군집 $i$에 속할 확률을 측정한 값이다. 식 (12)는 상대적 소속도의 제약식으로, 이 제약식으로 인하여 만일 입력 데이타에 오류가 있다면 Fuzzy Q-Map은 그 오류를 제외시키지 못하고 학습 결과에 반영하게 된다. 상대적 소속도가 이러한 단점을 가짐에도 불구하고 Fuzzy Q-Map에 이용한 이유는 상태공간 전역을 분석하고 일정한 범위의 행동 값을 예측하기 위해서이다.

단계 4: (행동의 예측)

Fuzzy Q-Map은 $\varepsilon$-*greedy* 정책에 따라서 상태 $s_t$에서 행동 $a_t$을 선택한다. $\varepsilon$-*greedy*는 상태공간을 탐험하기 위한 전략의 한 방법으로, $(1-\varepsilon)$%의 비율만큼 각 군집에서 Q 값이 가장 큰 행동을 선택하고 $\varepsilon$%의 비율만큼 랜덤하게 행동을 선택한다.

선택한 행동 $a_t$는 Fuzzy Q-Map이 전역적으로 제안하는 행동이고, 각 군집에서 제안하는 지역적인 행동 $a$들로부터 구한다. 행동 $a$은 각 군집에서 Q 값이 가장 큰 행동으로서 수식으로 정의하면 다음과 같다. $c_i$는 군집의 중심을 의미한다.

$$a = \max_{\arg a} q(c_i, a_j) \tag{13}$$

각 군집에서 $a$을 제안한다면, 소속도를 이용하여 최적의 행동 $a_t^*$을 다음 식 (14)와 같이 예측할 수 있다.

$$a_t^* = \frac{\sum_{i=1}^{c} m_{ik} \cdot a}{\sum_{i=1}^{c} m_{ik}} = \sum_{i}^{c} m_{ik} \cdot a\,,\ a_t = a_t^* \tag{14}$$

강화학습 문제의 행동 값은 이산적인 값이거나 일정한 범위 안의 실수 값이다. Fuzzy Q-Map이 제안하는 $a_t$는 소속도 값의 제약에 따라서 일정한 범위 안의 값이 된다.

단계 5: (행동 $a_t$의 수행)

에이전트는 선택한 행동 $a_t$을 수행하고, 환경으로부터 보상값 $r_{t+1}$을 받고, 행동 수행 결과로 전이한 새로운 상태 $s_{t+1}$을 인지한다.

단계 6: (Q 값의 갱신)

상태 $s_t$에서 선택한 행동 $a_t$을 평가하는 Q 값을 갱신한다. 갱신식은 식 1의 기본적인 Q-Learning의 갱신식을 이용한다. 상태 $s_t$에서 선택한 행동 $a_t$의 Q 값은 다음 식 (15)와 같이 Fuzzy Q-Map으로부터 구한다. 여기서 $q(c_i, a_t)$는 각 군집 $i$의 행동 $a_t$에 대한 지역적인 Q 값을 의미한다.

$$Q(s_t, a_t) = \sum_{i=1}^{c} (m_{it} \times q(c_i, a_t)) \quad (15)$$

행동 $a_t$를 수행한 결과로 도달한 상태인 $s_{t+1}$의 평가값 $f(s_{t+1})$은 다음과 같이 각 군집의 최대 Q 값과 소속도 $m_{i(t+1)}$로부터 구한다.

$$f(s_{t+1}) = \sum_{i=1}^{c} m_{i(t+1)} \times \max q(c_i, a) \quad (16)$$

위의 두 식을 기본적인 Q-Learning알고리즘의 갱신식에 대입하면 다음과 같다.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma f(s_{t+1}) - Q(s_t, a_t)) \quad (17)$$

$\alpha$는 학습률로서 0.5로 초기화되고 다음 식과 같이 $t$에 따라 감소한다.

$$\alpha = 0.5 \times 0.9^{\frac{t}{1000}} \quad (18)$$

단계 7: (승자 군집의 중심과 Q 값의 갱신)

입력 상태 $s_t$에 대하여, 소속도가 가장 높은 승자 군집(winner)의 중심 $c_w$과 지역적인 Q값을 갱신한다. Fuzzy Q-Map은 갱신하기 위해서 TD(Temporal Difference)에러와 소속도를 이용한다.

$$c_w \leftarrow c_w + (s_t - c_w) \times m_{wt} \times \alpha \quad (19)$$

$$q(c_w, a_t) \leftarrow q(c_w, a_t) + (Q(s_t, a_t) - q(c_w, a_t)) \times m \quad (20)$$

강화학습 알고리즘은 온 라인 학습의 특성 때문에, 훈련 데이타 집합과 소속도를 보존하는 것이 어려우므로 시간에 따른 차이와 소속도를 이용하여 갱신한다.

단계 8: (종료의 여부 검사)

종료 조건을 검사하여 만족하면 학습을 종료하고, 만족하지 않으면 단계 9로 간다. 종료 조건은 Fuzzy Q-Map의 군집의 중심과 Q 값에 변화가 없을 때일 수 있고, 또는 사용자가 지정한 반복의 횟수를 만족하도록 하는 방법이 있다. 본 논문에서는 반복 횟수를 사전에 지정하였다.

단계 9: (새로운 상태 $s_t$로 갱신하고 반복)

만일 $s_t$가 목표 상태라면 하나의 에피소드가 끝났으므로 상태 $s_t$를 랜덤하게 초기화한다. 그렇지 않다면 $s_t$를 $s_{t+1}$로 갱신한다. 단계 3으로 간다.

## 4. Fuzzy Q-Map의 성능 평가

### 4.1 마운틴 카 문제(Mountain Car Problem)

마운틴 카 문제는 그림 2와 같이 동력이 약한 자동차가 가파른 산길위로 목적지(goal)까지 운전하는 전략을 학습한다. 중력(gravity)은 자동차의 엔진보다 강하게 작용하는데, 만일 시작 위치가 가장 낮은 곳이라면 전속력으로 전진하더라도, 가파른 경사면을 올라가 목표에 도달할 수는 없다. 해결 방법은 먼저 목적지로부터 멀어져서 적절한 가속력을 얻은 후에 목적지로 전진하는 것이다. 자동차의 가능한 행동 $a_t$은 전진(+1), 후진(-1), 타성(0)이다. 자동차의 위치(position)를 $p_t$라 하고 속도(velocity)를 $v_t$라고 정의하면 다음 식 (21)에 의해서 두 변수는 갱신된다. bound 연산은 각 변수가 일정한 범위 안에 속하도록 한다. 두 변수의 범위는 다음 식 (22)와 같다. 위치 $p_{t+1}$가 오른쪽 한계이자 목적지인 0.5 이상이라면 하나의 에피소드는 끝나게 된다.

$$p_{t+1} = bound\ [p_t + v_{t+1}] \quad (21)$$

$$v_{t+1} = bound\ [v_t + 0.001 \times a_t - 0.0025\cos(3p_t)]$$

$$-1.2 \le p_{t+1} \le 0.5,\ -0.07 \le v_{t+1} \le 0.07 \quad (22)$$

마운틴 카 문제는 지연된 보상값(delayed reward)을 가진 문제로서 음수의 보상값을 주는 행동들이 장기적으로 보았을 때 최적임을 보여준다.

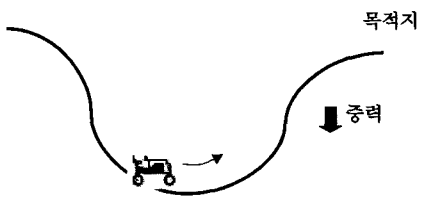



그림 2 마운틴 카 문제

### 4.2 실 험

강화학습에서 훈련 집합은 에이전트가 다수의 에피소드를 경험하는 과정에서 만나는 상태들로 구성된다. 그러므로 에이전트가 탐험하는 에피소드에 따라서 훈련 집합의 구성 상태들이 다르게 된다. 훈련 데이타는 학습을 위해 시스템으로 입력이 되고, 모든 단계들을 거치고 나면 시스템 밖으로 배출된다. 그러므로 훈련 데이타의 집합은 보존되는 것이 아니라 추상적인 개념이다. 본 실험에서는 보상함수는 다음 상태 $s_{t+1}$가 목표 상태라면 보상값을 1로 정하고, 그 밖의 상태는 -1로 정하였다. 학습 결과를 평가하기 위한 테스트 집합은 상태공간에서 고르게 분포된 100개의 상태들로 이루어져 있다.

4.2.1 새로운 훈련 데이타에 대한 Fuzzy Q-Map의 적응

Fuzzy Q-Map의 군집의 중심은 랜덤하게 초기화되

고, 훈련데이타에 적응하기 위해서 군집의 중심과 Q값은 계속해서 갱신된다. 다음 그림 3은 군집의 중심이 적응에 의하여 이동한 결과를 보인다. 오랜 시간동안 거대한 수의 훈련 데이타를 학습하면, 초기값과 상관없이 적응된 중심은 비슷한 좌표에 위치한다. 그림 3은 약 37만개의 상태들을 훈련했을 때의 각 군집의 중심을 보인다.

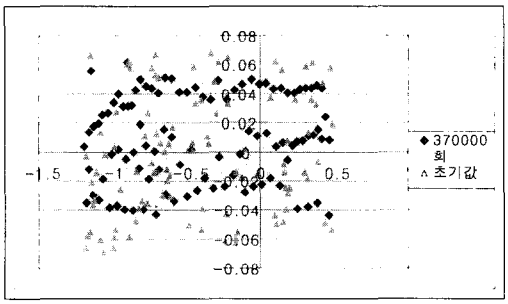



그림 3 군집의 중심 이동

4.2.2 비교실험

본 실험에서는 세 개의 알고리즘 즉 Fuzzy Q-Map과 이미 성능이 알려진 CMAC, 훈련 예제 집합을 이용한 LWR을 마운틴 카 문제에 적용하여 학습 속도와 예측률을 비교한다. CMAC와 LWR의 실험 결과는 William D. Smart의 논문[4]을 참조하였다.

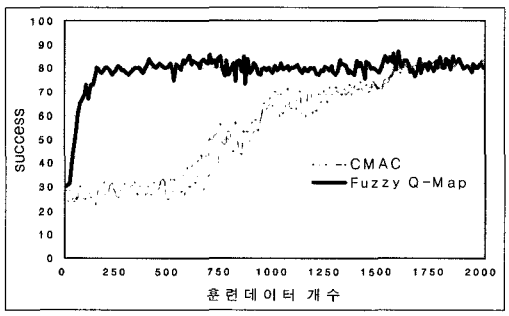



그림 4 CMAC와 Fuzzy Q-Map의 수렴 속도의 비교

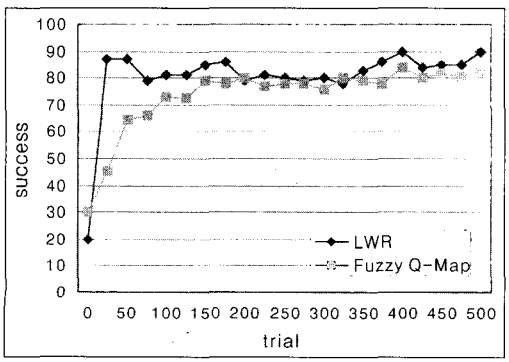



그림 5 LWR과 Fuzzy Q-Map의 수렴 속도의 비교

그림 4의 그래프에서, 사전 지식을 이용하지 않는 CMAC와 Fuzzy Q-Map의 학습 속도를 볼 수 있다. Fuzzy Q-Map은 CMAC보다 최고 예측률은 떨어지지만 학습 초기의 학습 속도는 가속화됨을 알 수 있다. 그림 5에서는 훈련 예제를 이용하는 LWR이 Fuzzy Q-Map보다 학습 속도나 예측률이 높다는 것을 보인다.

## 5. 결 론

강화학습은 환경과의 상호작용을 통하여 상태-행동쌍의 인과 관계를 경험하고 목표에 도달하기 위한 전략을 학습한다. 강화학습의 기본적인 알고리즘인 Q-learning은 거대한 상태공간을 갖는 문제인 경우, 메모리 문제와 학습 시간이 길다는 단점이 있다. 그리고 Q-learning에서 사용되는 보상함수는 다음상태가 목표인가에 따라 보상과 벌금을 주는 간단한 형태인데, 이러한 간단한 형태의 보상함수를 사용하면 학습 속도가 느리다. 만일 훈련 예제와 같은 유용한 사전지식이 있다면, 에이전트가 학습 초기에 상태공간을 탐험하는데 도움이 될 수 있고 학습 속도는 가속될 수 있다. 하지만 사전 지식은 전문가에 의하여 수집되는 것으로 쉽게 얻을 수 없는 경우도 있다.

본 논문에서는 사전지식이 충분히 주어지지 않고 거대한 상태공간을 갖는 문제를 강화학습 하기 위하여 온라인 퍼지 클러스터링을 기반으로 한 함수근사 방법인 Fuzzy Q-Map을 제안하였다. 제안한 Fuzzy Q-Map은 소속도 함수를 이용해서 경험하지 못한 새로운 훈련 데이타에 대해서도 유사한 군집들로부터 행동을 예측할 수 있고, 소속도에 따라서 갱신을 하므로 예측 에러를 감소시켜서 학습 속도를 가속화 한다. 또한 Fuzzy Q-Map은 군집의 중심과 Q값의 갱신을 유사한 군집에 대해서만 지역적으로 하므로 전체적인 전략을 바꿀 필요가 없게 된다. 지역적인 갱신은 간섭 현상을 줄일 수 있다.

현재 시뮬레이션 분야에서는 불확실성을 갖고 복잡한 시스템의 모델링하기 위하여 인공지능을 이용하는 추세이다. Fuzzy Q-Map은 시뮬레이션 소프트웨어로 사용될 수 있다.

향후 연구 과제는 다음과 같다. 첫째, 본 논문에서는 Fuzzy Q-Map을 마운틴 카 문제에 적용해 보았다. 그러나 제안한 알고리즘의 정당성을 입증하기 위해서는 실세계의 다양한 문제에 적용해 보아야 한다. 실세계의 문제는 시뮬레이션과 달라서 센서로부터 입력되는 값을 그대로 사용할 수 없다. Fuzzy Q-Map을 실세계에 적용하기 위해서 알고리즘의 수정이 필요하다. 둘째,

Fuzzy Q-Map에 맞는 적합도를 연구하여 학습 속도를 더욱 개선하겠다.

## 참 고 문 헌


[1] Richard Sutton, Andrew G. Barto, "Reinforcement Learning :An Introduction," MIT Press, 1998.

[2] Leslie Pack Kaelbling, Michael L. Littman, Andrew W. Moor, "Reinforcement Learning: A Survey," Journal of Artificial Intelligence Research, vol. 4, pp. 237-285, 1996.

[3] Pierre Yves Glorennce, "Reinforcement Learning : an Overview," Proceedings of the European Symposium on Intelligent Techniques, 2000.

[4] William Donald Smart, "Making Reinforcement Learning Work on Real Robots," Ph. D. Thesis, Brown University, 2002.

[5] A.K. Jain, M.N, Murty, P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, 1999.

[6] A. Baraldi, P.Blonda, "A survey of fuzzy clustering algorithms for pattern recognition," IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics), vol. 29, no. 6, pp. 778-785, 1999.

[7] Aristidis Likas, "A Reinforcement Learning Approach to On-line Clustering," Neural computation 11 (8): 1915-1932, 1999.

[8] Nicolas B. Karayiannis, James C. Bezdek, "An Integrated Approach to Fuzzy Learning Vector Quantization and Fuzzy c-Means Clstering," IEEE Transactions of Fuzzy systems, vol. 5, no. 4, 1997.

[9] 전종원, 민준영, "GLVQ클러스터링을 위한 필기체 숫자의 효율적인 특징추출 방법", 한국정보처리학회 논문지, vol. 2, no. 6, 1995.

[10] Barbara Hammer, Thomas Villmann, "Generalized Relevance Learning Vector Quantization," Neural Networks, vol. 15 no. 8-9, pp. 1059-1068, 2002.

[11] Shyn Jong Hu, "Pattern Recognition by LVQ and GLVQ Networks," http://neuron.et.ntust.edu.tw/homework/87/NN/87Homework%232/M8702043.

[12] Michael Herrmann, Ralf Der, "Efficient Q- Learning by Division of Labor," Proceedings of International Conference on Artificial Neural Networks, 1995.

[13] K. Yamada, M. Svinin, K. Ueda, "Reinforcement Learning with Autonomous State Space Construction using Unsupervised Clustering Method," Proceedings of the 5th International Symposium on Artificial Life and Robotics, 2000.

[14] Lionel Jouffe, "Fuzzy Inference System Learning by Reinforcement Methods," IEEE Transactions on Systems, Man and Cybernetics pp. 338-355, 1998.

[15] Andrea Bonarini, "Delayed Reinforcement, Fuzzy Q-Learning and Fuzzy Logic Controllers," In Herrera, F., Verdegay, J. L. (Eds.) Genetic Algorithms and Soft Computing, pp. 447-466, 1996.

[16] Pierre Yves Glorennec, Lionel Jouffe, "Fuzzy Q-Learning," Proceedings of Sixth IEEE International Conference on Fuzzy Systems, pp. 719-724, 1997.

[17] 정석일, 이연정, "분포기여도를 이용한 퍼지 Q-Learning", 퍼지 및 지능시스템 학회 논문지, vol. 11, no. 5, pp. 388-394, 2001.

[18] Richard S. Sutton, "Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding," Advances in Neural Information Processing Systems 8, pp. 1038-1044, MIT Press, 1996.

[19] R. Matthew Kretchmar, Charles W. Anderson, "Comparison of CMACs and Radial Basis Functions for Local Function Approximators in Reinforcement Learning," Proceedings of International Conference on Neural Networks, 1997.

[20] Juan Carlos Santamaria, Richard S. Sutton, Ashwin Ram, "Experiments with Reinforcement Learning in Problems with Continuous State and Action Spaces," COINS Technical Report 96-88, 1996.

[21] William D. Smart, Leslie Pack Kaelbling, "Practical Reinforcement Learning in Continuous Spaces," Proceedings of International Conference on Machine Learning, 2000.

[22] William D. Smart, Leslie Pack Kaelbling, "Reinforcement Learning for Robot Control," In Mobile Robots XVI, 2001.



이 영 아

1992년 동덕여자대학교 전자계산학과(학사). 1994년 동덕여자대학교 대학원 전자계산학과(공학석사). 2004년 8월 경희대학교 대학원 컴퓨터공학과(공학박사). 관심분야는 에이전트, 강화학습, 로보틱스



정 태 충

1980년 서울대학교 전자공학과(학사). 1982년 한국과학기술원 전자공학 전공(공학석사). 1987년 한국과학기술원 전자공학 전공(공학박사). 1987년~1988년 KIST 시스템 공학 센터 선임연구원. 1988년~현재 경희대학교 컴퓨터공학과 교수. 관심분야는 기계학습, 최적화, 에이전트, 보안