퍼지 클러스터링을 이용한 강화학습의 함수근사

이 영 아[†] · 정 경 숙[†] · 정 태 충[††]

요 약

강화학습을 적용하기에 적합한 많은 실세계의 제어 문제들은 연속적인 상태 또는 행동(continuous states or actions)을 갖는다. 연속 값을 갖는 문제인 경우, 상태공간의 크기가 거대해져서 모든 상태-행동 쌍을 학습하는데 메모리와 시간상의 문제가 있다. 이를 해결하기 위하여 학습된 유사한 상태로부터 새로운 상태에 대한 추측을 하는 함수 근사 방법이 필요하다. 본 논문에서는 1-step Q-learning의 함수 근사를 위하여 퍼지 클러스터링을 기초로 한 Fuzzy Q-Map을 제안한다. Fuzzy Q-Map은 데이터에 대한 각 클러스터의 소속도(membership degree)를 이용하여 유사한 상태들을 군집하고 행동을 선택하고 Q값을 참조했다. 또한 승자(winner)가 되는 퍼지 클러스터의 중심과 Q값은 소속도와 TD(Temporal Difference) 에러를 이용하여 갱신하였다. 본 논문에서 제안한 방법은 마운틴 카 문제에 적용한 결과, 빠른 수렴 결과를 보였다.

Function Approximation for Reinforcement Learning using Fuzzy Clustering

Young Ah Lee[†] · Kyoung Sook Jung[†] · TaeChoong Chung[††]

ABSTRACT

Many real world control problems have continuous states and actions. When the state space is continuous, the reinforcement learning problems involve very large state space and suffer from memory and time for learning all individual state-action values. These problems need function approximators that reason action about new state from previously experienced states. We introduce Fuzzy Q-Map that is a function approximators for 1 - step Q-learning and is based on fuzzy clustering. Fuzzy Q-Map groups similar states and chooses an action and refers Q value according to membership degree. The centroid and Q value of winner cluster is updated using membership degree and TD(Temporal Difference) error. We applied Fuzzy Q-Map to the mountain car problem and acquired accelerated learning speed.

키워드 : 강화학습(Reinforcement Learning), Q-learning, 함수근사(Function Approximation), Fuzzy Q-Learning, 퍼지 클러스터링(fuzzy clustering), 소속도(membership degree)

1. 서 론

교사학습(supervised learning)은 목표값이 주어지는 예제의 집합을 학습하는 반면, 강화학습은 입력으로 들어온 상태에서 행동을 선택했을 때, 다음 상태가 목표(goal)인가의 여부에 따라 주어지는 보상값(reward)과 현재까지 경험해서 얻은 불완전한 지식을 기반으로 계속 반복해서 학습해야 한다. 학습 속도는 느리지만 강화학습은 변화하는 개념과 온라인 학습이 가능한 방법이다[1].

강화학습을 적용할 수 있는 실세계의 문제들은 연속된 상태와 행동들(continuous states and actions)을 갖는 경우가 많다. 기본적인 강화학습 알고리즘들은 이산 상태와 이산 행동들(discrete states and actions)을 다루기 때문에 실세계의 문제에 그대로 적용할 수 없다[2, 3]. 이 문제를 해결하기 위한 방법으로 연속적인 상태공간과 행동을 이산화 할 수 있는데, 매우 미세하게 이산화(quantization)를 하면 상태공간이 거대해져서 최적의 학습 결과를 얻기까지 많은 시간과 기억장소를 필요로 하게 된다. 기본 강화학습 알고리즘인 Q-learning에서 학습 결과를 기억하기 위해 사용하는 lookup table은 상태공간의 크기에 따라 거대해지고, 모든 이산 상태-행동 쌍의 평가 값인 Q값을 저장하므로 학습 내용을 모두 담고 있지만 단위 정보이고 정리된 내용은 아니다. 또한 Q-learning은 상태-행동 쌍(state-action pair)의 평가값인 Q값을 갱신하기 위해서, 다른 유사한 상태-행동 쌍들의 학습 내용은 참조하지 않고 해당 상태-행동 쌍의 보상 값과 Q값만을 이용하므로 학습 속도가 느리다. 이러한 문제들을 해결하기 위하여 유사한 상태들을 일반화 시키는 함수근사(function approximation) 방법을 이용한다. 많은 연구에서 함수근사 방법으로서 전 방향 또는 역방향 신경망(Feedforward or Backpropagation Neural Network)[5], 자기

† 준 회 원 : 경희대학교 대학원 컴퓨터공학과
†† 정 회 원 : 경희대학교 컴퓨터공학과 교수
논문접수 : 2003년 7월 16일, 심사완료 : 2003년 8월 27일

형상화 지도(Self-Organizing features Map : SOM), CMAC (Cerebella Model Articulation Controller)[1]를 이용하였고, Q-learning을 퍼지 환경으로 확장하는 연구[5-9]가 있었다. 신경망은 블랙박스이므로 행동 선택에 대한 설명이 불가능하고, SOM의 이웃개념(neighborhood concept)은 상태-행동별 전략학습을 하는 Q-learning에 적용하기가 어렵다. Glorennce[7]는 기본적인 강화학습 알고리즘이 이산의 상태와 행동을 다루는 한계를 극복하기 위해서 Q-learning과 FIS (Fuzzy Inference System)를 접목한 FQL(Fuzzy Q-Learning)과 FACL(Fuzzy Actor Critic Learning) 알고리즘[7, 8]을 연구하였다. FQL과 FACL은 퍼지 규칙(fuzzy rule)의 조건부는 사전처리로 고정시키고 결론부를 강화학습을 통해 조정하는 방법이다.

본 논문에서는 퍼지 클러스터링의 소속도(membership degree)를 이용하여 Q-함수를 근사화하는 Fuzzy Q-Map을 제안한다. 실세계의 많은 제어문제들은 상태와 행동이 연속값(continuous value)으로 표현되고, 훈련데이터에 대한 분류정보 또는 분포가 사전에 주어지지 않으며, 유사한 데이터 집합사이의 경계가 명확하지 않으므로 함수 근사방법으로 비교사 학습 방법(unsupervised learning)인 퍼지 클러스터링이 적합하다고 보았다[10]. Fuzzy Q-Map에서는 입력으로 들어온 상태가 각 클러스터에 속하는 정도를 나타내는 소속도를 이용하여 여러 퍼지 클러스터에 속하게 된다. 각 클러스터가 제안하는 행동은 소속도 만큼 참조되고, 승자(winner)와 유사한 클러스터들의 현재까지의 학습결과를 바탕으로 승자의 중심(centroid)과 Q값은 새로운 입력에 적응된다. Fuzzy Q-Map은 상태공간을 퍼지 변수로 코딩하는 사전처리가 필요 없으므로 환경의 변화와 새로운 데이터에 대한 적응력을 갖고 있다. 또한 입력과 거리가 가장 가까운 클러스터뿐만 아니라 유사한 여러 클러스터가 제안하는 행동과 Q값을 참조하므로 학습속도가 가속된다.

본 논문의 구성은 다음과 같다. 2장에서는 Q-learning 알고리즘을 요약하고, Fuzzy Q-Map의 기초가 된 FQL(Fuzzy Q-Learning)과 FCM(Fuzzy C means)알고리즘을 살펴본다. 3장에서는 본 논문에서 제안한 Fuzzy Q-Map에 대하여 설명한다. 4장에서는 마운틴 카 문제(mountain car problem)를 Fuzzy Q-Map을 이용하여 학습한 결과를 분석하였다. 5장에서는 결론 및 향후 연구 과제를 제시하였다.

2. 관련 연구

2.1 Q-learning

Watkins가 제안한 Q-learning[1]은 강화학습의 대표적인 알고리즘이다. Q-learning은 현재 상태 s_t에서 어떤 행동 a_t를 수행하였을 때 받은 강화값에 대한 근사값을 상태-행동쌍에 대한 Q함수 $Q(s_t, a_t)$에 할당한다. 그리고 다음 상태 s_{t+1}에서 Q함수 $Q(s_{t+1}, a_{t+1})$가 최대가 되는 행동 a_{t+1}을 선택하여 현재 상태의 Q함수 값을 다음 식 (1)과 같이 갱신한다.

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \quad (1)$$

식 (1)을 이용한 Q-learning을 1-step Q-learning이라 한다. 1-step Q-learning에서는 Q값을 저장하기 위하여 look-up table을 사용하였고, Q함수가 수렴하면 최적의 정책은 각 상태에서 가장 큰 Q-값을 갖는 행동을 선택함으로써 구현된다.

2.2 FQL(Fuzzy Q-Learning)

FQL[7-9]은 Glorennce가 Watkins의 Q-learning을 바탕으로 개발한 학습방법이다. FQL은 Q함수를 근사화하기 위하여 FIS(Fuzzy Inference System)를 이용하는데, 퍼지 규칙(fuzzy rule)의 결론부를 강화학습의 보상값을 이용하여 조정하고, 조건부는 사전 정보에 의하여 상태공간을 퍼지 입력변수로 코딩한다. 각 규칙은 행동들의 경쟁을 위하여 각 행동의 Q값을 관리한다. 규칙의 형태는 다음과 같다. 규칙 i에서 가능한 행동 $a[i, j]$들이 J개이고 각각 Q값 $q[i, j]$를 기억한다. S_i는 퍼지 레이블(fuzzy label)이다.

if x is S_i *then* $a[i, 1]$ *with* $q[i, 1]$
or $a[i, 2]$ *with* $q[i, 2]$
.....
or $a[i, J]$ *with* $q[i, J]$

함수 $x \rightarrow a_i(x)$은 입력 상태 x의 규칙 i에 대한 진리값으로 [0, 1] 사이의 값을 갖는다. FIS의 N개의 규칙이 추론하는 행동 $a(x)$은 다음 식 (2)와 같다.

$$a(x) = \frac{\sum_{i=1}^{N} a_i(x) \times a_i}{\sum_{i=1}^{N} a_i(x)} \quad (2)$$

추론된 행동 a의 Q값은 다음 식 (3)과 같다. i^+은 EEP (Exploration/Exploitation Policy)에 의하여 규칙 i에서 선택된 행동이고, $q[i, i^+]$은 규칙 i에서 선택된 행동 i^+의 Q값을 표시한다.

$$Q(x, a) = \frac{\sum_{i=1}^{N} a_i(x) \times q[i, i^+]}{\sum_{i=1}^{N} a_i(x)} \quad (3)$$

상태 x에 대한 평가값은 다음 식 (4)와 같다. i^*은 규칙 i에서 최대의 Q값을 갖는 행동이다.

$$V(x)=\frac{\sum_{i=1}^{N} a_i(x)\times q[i,\ i^*]}{\sum_{i=1}^{N} a_i(x)} \tag{4}$$

Q값은 식 (1)~식 (4)를 이용하여 갱신된다. 식 (5)에서 ΔQ는 $\Delta Q=r+\gamma V(y)-Q(x,\ a)$로서 에러를 의미하며, ε은 학습율이다.

$$\Delta q[i,\ i^+]=\varepsilon\times\Delta Q\frac{a_i(x)}{\sum_{i=1}^{N} a_i(x)} \tag{5}$$

FQL에서는 학습 속도를 높이기 위해서 행동의 적합도(eligibility) $e[i,\ j]$을 다음과 같이 정의하여 식 (5)를 수정하였다.

$$e[i,\ j]=\begin{cases}\lambda\gamma e[i,\ j]+\dfrac{a_i(x)}{\sum_{i=1}^{N} a_i(x)} & \text{if } j=i^+\\ \lambda\gamma e[i,\ j] & elsewhe\end{cases} \tag{6}$$

$$\Delta q[i,\ i^+]=\varepsilon\times\Delta Q\times e[i,\ j] \tag{7}$$

2.3 FCM(Fuzzy C Means)알고리즘

FCM 알고리즘[4, 10]은 $u_k(k=1,\ 2,\ \cdots,\ K)$로 정의된 K개의 데이터를 c개의 퍼지 클러스터로 분할하고, 식 (8)의 목적함수(objective function) J를 최소화하도록 클러스터의 중심을 찾는다. FCM은 K-means 알고리즘에 분할의 fuzziness를 증감시키는 파라미터 $q\in(1,\ INF)$를 추가한 것으로, 각 데이터가 여러 클러스터에 소속될 수 있다. 식 (8)에서 M은 훈련 집합의 각 데이터가 각 클러스터에 속하는 소속정도를 저장하는 $c\times K$ 크기의 행렬이다.

$$J(M,\ c_1,\ c_2,\ \cdots,\ c_c)=\sum_{i=1}^{c}\sum_{k=1}^{K} m_{ik}^q\ d_{ik}^2 \tag{8}$$

M : 소속도 행렬 q : fuzziness c_i : 퍼지 클러스터 i의 중심

$d_{ik}=\|u_k-c_i\|$: 데이터 u_k로부터 클러스터 i의 중심 c_i까지의 유클리드 거리

m_{ik} : 클러스터 i에 대한 데이터 u_k의 소속도

목적함수 J가 최저한도에 도달하기 위해서 다음 식 (9)와 식 (10)을 만족하여야 한다. 소속도 m_{ik}은 [0, 1] 사이의 값을 갖고, 각 데이터에 대한 모든 클러스터 소속도의 합은 1이다. 각 클러스터의 중심은 식 (10)에 의하여 갱신된다.

$$m_{ik}=\frac{1}{\sum_{j=1}^{c}\left(\dfrac{d_{ik}}{d_{jk}}\right)^{2/(q-1)}},\qquad \sum_{j=1}^{c} m_{ji}=1 \tag{9}$$

$$c_i=\frac{\sum_{k=1}^{K} m_{ik}^q\ u_k}{\sum_{k=1}^{K} m_{ik}^q} \tag{10}$$

FCM은 각 퍼지 클러스터의 중심을 갱신하기 위해서 K개 데이터의 소속도를 저장한 행렬 M을 관리해야 한다. 그러나 Q-learning에 사용되는 훈련 데이터 집합은 상태공간의 크기에 비례해서 커지므로 lookup table과 마찬가지로 소속도 행렬 M을 관리하기 어렵다. 본 논문에서 제안하는 Fuzzy Q-Map은 클러스터의 중심과 각 행동의 Q값을 기억하는 구조로서 승자 클러스터의 중심과 Q값은 입력과의 오류와 소속도를 이용하여 적응시켰다.

3. Fuzzy Q-Map

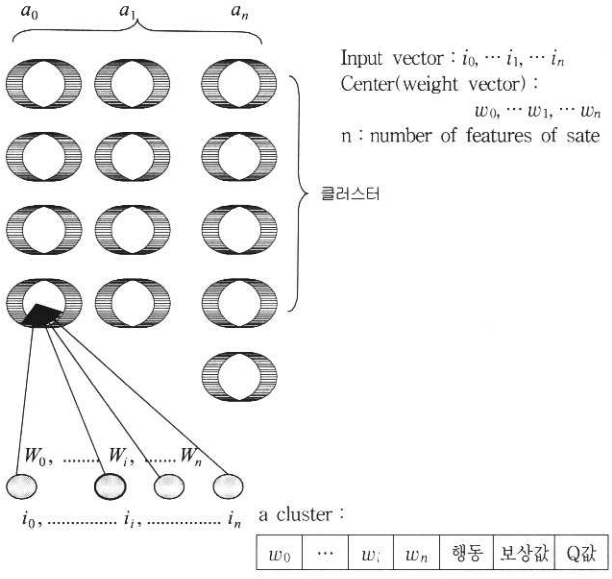
본 논문에서 제안한 Fuzzy Q-Map은 1-step Q-learning의 함수 근사를 위하여 FCM의 소속도를 이용한다. 상태공간을 구성하는 요소들(features)과 행동들이 연속값을 갖는 경우, 상태공간은 거대해지므로 Q-learning의 lookup table은 사용불가능하고, 식 (1)에서 보여주듯이 현재 경험한 상태-행동 쌍의 Q값과 보상값만을 이용해서 전략을 학습하므로 학습속도가 느리다. Fuzzy Q-Map은 소속도를 이용하여 유사한 상태들을 군집하여 lookup table의 크기를 줄이고, 승자 퍼지 클러스터와 가까운 퍼지 클러스터들이 제안하는 행동들도 소속도 만큼 참조하여 학습하므로 학습 속도가 빨라진다.

Fuzzy Q-Map은 각 퍼지 클러스터의 중심과 행동들의 Q값을 기억하는 구조이다. 상태 공간의 분할을 선행처리 하는 경우에는 도메인을 잘 아는 사용자의 사전지식이 필요하고 분할이 고정 된다. 반면 Fuzzy Q-Map은 온라인으로 Q-learning을 진행하면서 각 퍼지 클러스터의 중심 값을 새로운 경험에 적응시키므로 사전처리가 필요 없고, 분할은 조금씩 변하게 된다.

Fuzzy Q-Map에서 각 클러스터의 중심 값은 규칙의 조건부에 해당하는데, 유사한 상태들을 대표한다. 최종 학습된 결과를 담고 있는 Fuzzy Q-Map은 상태공간을 지역적으로 분석한 결과이고, 별도의 복잡한 규칙 추출과정이 필요 없다.

Fuzzy Q-Map은 (그림 1)과 같이 2차원으로, 행의 개수는 사용자가 지정한 클러스터의 개수이고, 열의 개수는 상태공간에서 가능한 행동의 수이다. 그리고 강화학습을 하는 동안 여러 에피소드들(임의의 상태에서 출발하여 목표 상태에 도달할 때까지의 경로)을 경험해야 하는데, 목표상태(goal state)는 한 에피소드의 끝으로서 다음 상태로 전이되지 않으므로 군집된 다른 상태와 같이 다룰 수 없다. 그러므로 목표상태를 중심으로 갖는 클러스터를 추가하여, Fuzzy Q-Map을 구성하는 노드의 수는 (클러스터의 수×가능한 행동의 수) + 1이 된다.

각 퍼지 클러스터는 (그림 1)과 같이 중심과 행동, 그 행동을 수행했을 때 받은 보상값, t시점에서의 Q값을 기억하고 있다. 상태는 n차원 벡터이고, 현재 상태 $i_0, i_1, \cdots, i_n$와 각 퍼지 클러스터의 중심 $w_0, w_1, \cdots, w_n$과의 유클리드 거리를 측정하여 소속도를 구한다. 소속도는 행동의 선택, Q값 참조, 클러스터의 중심과 Q값의 갱신에 사용된다.



(그림 1) Fuzzy Q-Map과 클러스터의 구조

제안한 Fuzzy Q-Map을 이용한 Q-learning을 정리하면 다음과 같다.

단계 1 : Fuzzy Q-Map의 각 클러스터 중심을 랜덤하게 초기화한다. 각 클러스터의 보상값과 Q값은 0으로 초기화한다.

단계 2 : 입력 상태 u_t를 랜덤하게 초기화한다. t는 하나의 상태를 처리하는 시점을 표시하며 지금까지 훈련에 사용된 상태의 개수이기도 하다.
훈련 데이터 집합은 에피소드들의 집합으로서 상태공간을 탐험하는 과정에서 만나게 되는 상태들로 이루어진다. u_t는 한 에피소드를 구성하는 t 시점의 입력 상태로서 질의(query)이다.

단계 3 : u_t가 각 클러스터 i에 속하는 소속도 m_{it}를 구한다. u_t에 대한 $m_{it}(1 \le i \le c)$의 합은 1이다. q는 fuzziness의 파라미터로서 사용자가 입력한다.
소속도는 입력상태와 가장 가까운 승자뿐만 아니라 거리가 가까운 클러스터들을 알 수 있게 한다.

$$m_{it} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{it}}{d_{jt}} \right)^{2/(q-1)}}, \qquad \sum_{i=1}^{c} m_{it} = 1 \qquad (11)$$

단계 4 : 상태 u_t에서 $\varepsilon - greedy$ 전략을 이용하여 행동 a_t을 선택 한다.
최적의 행동 a_t^*은 다음과 같이 식 (12)에 의하여 계산한다. 각 클러스터 i가 제안하는 행동 a_t^i을 소속도 m_{it}만큼 참조한다. 즉 거리가 가까운 클러스터들의 제안이 행동 선택에 많은 영향을 주게 된다. a_t^i는 각 클러스터에서 가능한 행동 중 Q값이 가장 큰 행동이다. c_i는 각 클러스터의 중심이다.

$$a_t^* = \frac{\sum_{i=1}^{c} m_{ik} \cdot a_t^i}{\sum_{i=1}^{c} m_{ik}} = \sum_{i}^{c} m_{ik} \cdot a_t^i \qquad (12)$$

$$a_t^i = \max_{arg a} Q(c_i, a), \quad a \in A \quad A : action\ set$$

u_t에서 선택한 행동 a_t의 평가값인 Q값은 다음과 같이 계산된다.

$$Q(u_t, a_t) = \sum_{i=1}^{c} m_{it} \cdot qval(a_t) \qquad (13)$$

$qval(a_t)$: 각 클러스터의 행동 a_t에 대한 Q값

상태 u_t에서 최대의 Q값은 다음과 같이 각 클러스터의 최대 Q값과 소속도로부터 계산한다.

$$f(u_t) = \sum_{i=1}^{c} m_{it} \cdot \max Q(c_i, a), \quad a \in A \qquad (14)$$

단계 5 : 선택한 행동 a_t를 수행한 후, 환경으로부터 다음 상태 u_{t+1}와 보상값 r_{t+1}를 얻는다. 식 (15)를 이용하여 소속도가 가장 높은 승자 클러스터의 Q값을 갱신한다. β는 Fuzzy Q-Map의 학습률로서 0.5로 초기화되고, 식 (18)과 같이 t에 따라 감소한다.

$$Q(u_t, a_t) = Q(u_t, a_t) + \beta(r_{t+1} + \gamma f(u_{t+1}) - Q(u_t, a_t)) \qquad (15)$$

단계 6 : 입력데이터 u_t에 대한 소속도가 가장 높은 Fuzzy Q-Map의 승자 클러스터의 중심 c_w^t와 Q값을 갱신한다. FCM은 소속도 행렬을 이용하여 각 클러스터의 중심을 갱신하지만 Fuzzy Q-Map은 새로운 입력과 기존의 값의 에러와 소속도를 이용하여 승자 클러스터의 중심을 갱신한다.

$$c_w^t = c_w^t + (u_t - c_w^t) \cdot m_{wt} \cdot \beta \qquad (16)$$

$$Q^t(c_w^t, a_t) = Q^t(c_w^t, a_t) + (Q^t(c_w^t, a_t) - Q(u_t, a_t)) \cdot m_{wt} \qquad (17)$$

$$\beta = 0.5 \cdot 0.9^{\frac{t}{000}} \qquad (18)$$

단계 7 : 만일 $u_t + 1$이 목표상태라면 상태 u_t를 랜덤하게 초

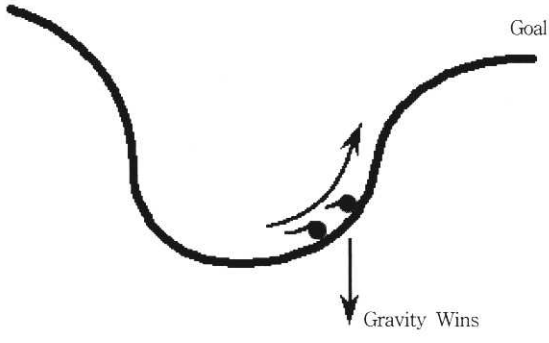
기화한다. 그렇지 않다면 u_t를 u_{t+1}로 갱신한다.

단계 8 : 종료조건을 만족하지 않으면 단계 3으로 간다.

4. 실험과 분석

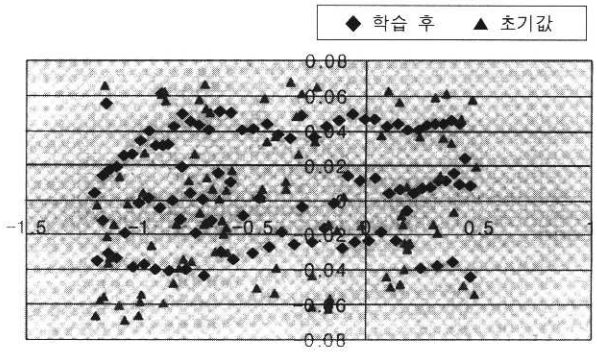
본 논문에서 제안한 Fuzzy Q-Map을 평가하기 위하여, 연속적인 상태공간을 갖는 마운틴 카 문제에 적용하였다.

마운틴 카 문제는 (그림 2)와 같이 자동차가 목표에 도달하기 위하여 적절한 가속을 얻도록 관성(coast), 전진(forward), 후진(backward)의 행동을 학습하는 것이다. 마운틴 카 문제에서 상태공간을 구성하는 요소는 위치(P)와 속도(V)이고, 각각 일정한 범위에 속하는 연속값을 갖는다. 위치(P)는 $-1.2 \le P \le 0.5$ 범위 내의 값이고, 속도(V)는 $-0.07 \le 0.07$ 사이의 값을 갖는다. 자동차는 동력이 약한 엔진과 중력의 제약조건을 이기고 목표지점인 0.5에 도달하여야 한다.



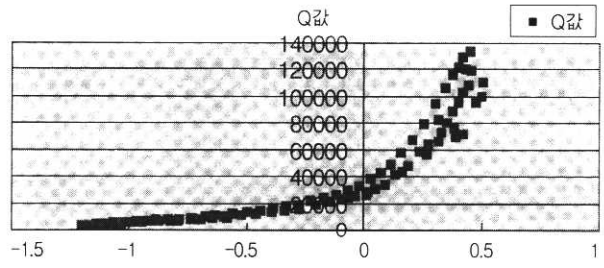
(그림 2) 마운틴 카 문제

본 실험에서 Fuzzy Q-Map은 301개의 노드(100개의 클러스터×3개의 이산 행동+1)로 이루어져 있고, fuzziness는 2로 지정하였다.



(그림 3) 클러스터 중심의 이동

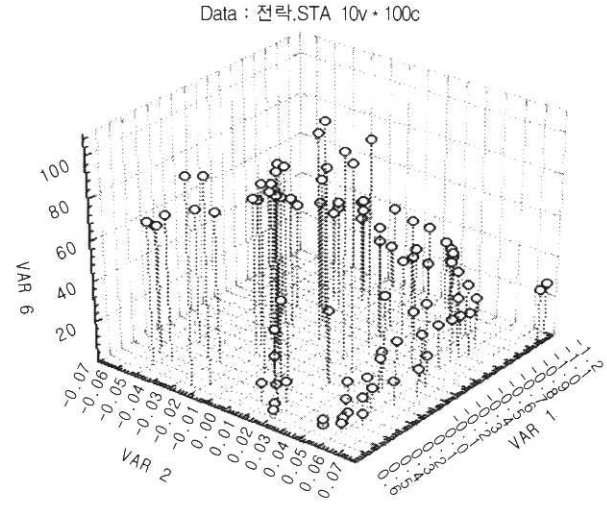
(그림 3)은 중심을 랜덤하게 초기화하고, 목표에 도달한 횟수가 5000번이 될 때까지 약 37만개의 상태들을 이용하여 훈련한 후, 이동한 클러스터의 중심을 나타낸다. 중심을 다르게 초기화해도 학습이 끝난 후의 적응된 중심들은 비슷한 좌표였다.



(그림 4) Fuzzy Q-Map의 위치별 Q값

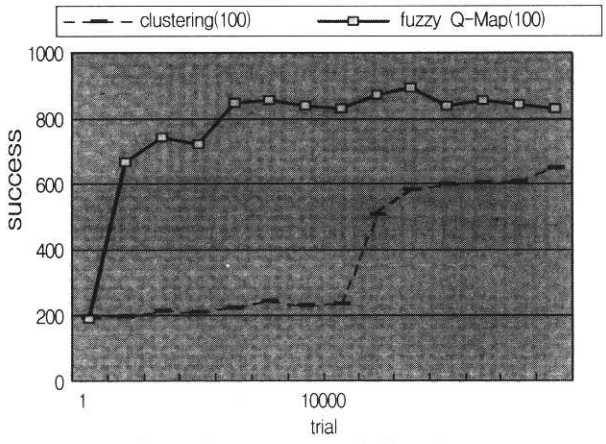
(그림 4)는 학습이 끝난 후, Fuzzy Q-Map의 각 클러스터 중심에서의 Q값을 그래프로 그린 것이다. 목표(0.5)에 가까울수록 Q값은 커지므로 목적을 쉽게 이룰 수 있음을 의미한다.

(그림 5)는 임의의 위치(VAR 1)와 임의의 속도(VAR2)에서 출발하여, 목표에 도달하기 위해 필요한 이동 횟수를 나타낸다. 시작 위치가 목표에 가깝고 속도가 0.01이상이면 적은 수의 이동으로 목표에 도달할 수 있고, 또한 목표에서 멀지만 속도가 양수이면 쉽게 목표에 도달 할 수 있다. 그러나 위치와 상관없이 속도가 0미만이면 목표에 도달하기 위한 가속을 얻어야 하므로 후진이 필요하여 이동수가 많아짐을 보인다.



(그림 5) Fuzzy Q-Map을 이용한 이동횟수

(그림 6)은 Fuzzy Q-Map의 소속도가 Q-learning의 학습을 가속시킴을 보인다. 그래프에서 X축은 훈련에 사용된 상태-행동 쌍의 개수를 표시하고, Y축은 상태 1000개의 테스트 집합에서 목표에 도달한 상태의 수를 나타낸다. 비교대상은 Fuzzy Q-Map의 소속도를 적용하지 않고, 유클리드 거리가 가까운 유사한 클러스터들을 클러스터링한다. 질의 상태는 거리가 가장 가까운 하나의 승자 클러스터에 속하게 되고, 승자 클러스터의 행동을 선택하도록 하였다. 클러스터링의 그래프는 10000개 이상의 시도에서도 약 60%의 성공률을 보인다. 그러나 소속도를 이용한 Fuzzy Q-Map은 5000번의 시도이후에 최고의 성능에 가까워짐을 보인다.



(그림 6) Fuzzy Q-Map의 학습 속도

5. 결론 및 향후 연구 과제

강화학습은 모델이 알려지지 않은 환경과 상호작용을 하면서 제어 규칙을 학습하는 방법으로, 실세계의 제어문제 학습에 적합하다. 그러나 실세계의 많은 문제들은 연속적인 상태와 행동을 가지므로, 이산데이터를 다루는 Q-learning 알고리즘을 그대로 이용할 수 없다. 본 논문에서는 사전에 데이터의 분포가 주어지지 않고, 보상값만을 이용하는 강화학습의 함수 근사방법으로서 비교사 학습 방법인 퍼지 클러스터링이 적합하다고 보았고, 퍼지 클러스터링 알고리즘인 FCM을 기초로 한 Fuzzy Q-Map을 제안했다. Fuzzy Q-Map은 소속도를 이용하여 각 퍼지 클러스터가 제안하는 행동을 조합하여 선택하였고, 각 클러스터의 중심과 Q값의 갱신에 TD 에러와 소속도를 이용하였다. Fuzzy Q-Map을 마운틴 카 문제에 적용한 결과 학습 속도의 개선과 새로운 훈련 집합에 적응력이 있음을 알 수 있었다.

그러나 Fuzzy Q-Map은 사용자가 클러스터의 개수를 지정해야 하고, 학습 결과 유사한 전략의 클러스터들이 존재하였다. 향후 중복된 클러스터들을 제거하는 방법과 성능 평가에 대한 연구가 필요하다.

참 고 문 헌

[1] Richard S. Sutton and Andrew G. Barto. "*Reinforcement Learning : An Introduction*," The MIT Press, Cambridge, MA., 1998.

[2] Stephan ten Hagen and Ben Krose, "Q-learning for System with continuous state and action spaces," BENELEARN 2000, 10th Belgian-Dutch conference on Machine Learning.

[3] Chris Gaskett, David Wettergreen, and Alexander Zelinsky, "Q-learning in continuous state and action spaces," Australian Joint Conference on Artificial Intelligence 1999.

[4] Jan Jantzen, "Neurofuzzy Modelling," Technical Report 98-H-869 (soc), Technical University of Denmark : Dept. of Automation, http://fuzzy.iau.dtu.dk/download/soc.pdf, 1998. Lecture notes, pp. 14.

[5] 전효병, 이동욱, 김대준, 심귀보, "퍼지추론에 위한 리커런트 뉴럴 네트워크 강화학습", 한국퍼지및지능시스템학회 '97년도 춘계학술대회논문집, 1997.

[6] 정석일, 이연정, "분포 기여도를 이용한 퍼지 Q-learning", 퍼지및지능시스템학회논문지, Vol.11, No.5 pp.388-394, 2001.

[7] Pierre Yves Glorennec, Lionel Jouffe, "Fuzzy Q-learning," Proceedings of Fuzz-Ieee'97, Sixth International Conference on Fuzzy Systems, Barcelona, pp.719-724, July, 1997.

[8] Lionel Jouffe, "Fuzzy Inference System Learning by Reinforcement Methods," Ieee Transactions on System, Man and Cybernetics, Vol.98, No.3, August, 1998.

[9] Andrea Bonarini, "Delayed Reinforcement, Fuzzy Q-learning and Fuzzy Logic Controllers," In Herrera, F., Verdegay, J. L. (Eds.) Genetic Algorithms and Soft Computing, (Studies in Fuzziness, 8), Physica-Verlag, Berlin, D., pp.447-466.

[10] Artistidis Likas, "A Reinforcement Learning Approach to On-Line Clustering," Neural Computation, Vol.11, No.8, pp. 1915-1932, 1999.

이 영 아

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

1992년 동덕여자대학교 전자계산학과(학사)

1994년 동덕여자대학교 대학원 전자계산학과(공학석사)

1999년~현재 경희대학교 대학원 컴퓨터공학과 박사수료

관심 분야 : 강화학습, 에이전트, 데이터마이닝, 로보틱스

정 경 숙

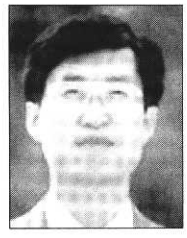
e-mail : jungks@iislab.kyunghee.ac.kr

1995년 경희대학교 수학과 졸업

1997년 경희대학교 컴퓨터공학과 석사

1999년~현재 경희대학교 대학원 컴퓨터공학과 박사수료

관심분야 : 정보보호, 인공지능, 전자상거래, 기계학습, 에이전트



정 태 충

e-mail : tcchung@khu.ac.kr

1980년 서울대학교 전자공학과(학사)

1982년 한국 과학 기술원 전자공학전공(공학석사)

1987년 한국 과학 기술원 전자공학전공(공학박사)

1987년~1988년 KIST 시스템 공학센터 선임연구원

1988년~현재 경희대학교 컴퓨터공학과 교수

관심분야 : 기계학습, 보안, 최적화, 에이전트