

퍼지 추론 기반의 멀티에이전트 강화학습 모델

Multi-Agent Reinforcement Learning Model based on Fuzzy Inference

이봉근*, 정재두**, 류근호*
충북대학교 전기전자 컴퓨터공학부*, 국방부 전자계산소**

Bong Keun Lee(bong9065@hanmail.net)*, Jae Du Chung(chungjaedu@hanmail.net)**,
Keun Ho Ryu(khryu@dblab.chungbuk.ac.kr)*

요약

강화학습은 최적의 행동정책을 구하는 최적화 문제로 주어진 환경과의 상호작용을 통해 받는 보상 값을 최대화하는 것이 목표이다. 특히 단일 에이전트에 비해 상태공간과 행동공간이 매우 커지는 다중 에이전트 시스템인 경우 효과적인 강화학습을 위해서는 적절한 행동 선택 전략이 마련되어야 한다. 본 논문에서는 멀티에이전트의 효과적인 행동 선택과 학습의 수렴속도를 개선하기 위하여 퍼지 추론 기반의 멀티에이전트 강화학습 모델을 제안하였다. 멀티 에이전트 강화학습의 대표적인 환경인 로보컵 Keepaway를 테스트 베드로 삼아 다양한 비교 실험을 전개하여 에이전트의 효율적인 행동 선택 전략을 확인하였다. 제안된 퍼지 추론 기반의 멀티에이전트 강화학습모델은 다양한 지능형 멀티 에이전트의 학습에서 행동 선택의 효율성 평가와 로봇축구 시스템의 전략 및 전술에 적용이 가능하다.

■ 중심어 : | 멀티에이전트 | 강화학습 |

Abstract

Reinforcement learning is a sub area of machine learning concerned with how an agent ought to take actions in an environment so as to maximize some notion of long-term reward. In the case of multi-agent, especially, which state space and action space gets very enormous in compared to single agent, so it needs to take most effective measure available select the action strategy for effective reinforcement learning. This paper proposes a multi-agent reinforcement learning model based on fuzzy inference system in order to improve learning collect speed and select an effective action in multi-agent. This paper verifies an effective action select strategy through evaluation tests based on Robocup Keepaway which is one of useful test-beds for multi-agent. Our proposed model can apply to evaluate efficiency of the various intelligent multi-agents and also can apply to strategy and tactics of robot soccer system.

■ keyword : | Multi-Agent | Reinforcement Learning |

1. 서론

지능형 로봇이나 가상의 캐릭터 연구에서 에이전트

가 현재 상황을 인식하여 상황에 맞는 적절한 행동을 선택하는 행동 선택 문제에 대한 많은 연구가 진행되어

* 이 논문은 2009년 교육과학기술부(지역거점연구단육성사업/충북 BIT 연구중심대학육성사업단)의 지원을 받아 수행된 연구임

* This research was financially supported by the Ministry of Education, Science Technology (MEST) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Regional Innovation

접수번호 : #090918-002

심사완료일 : 2009년 09월 25일

접수일자 : 2009년 09월 18일

교신저자 : 정재두, e-mail : chungjaedu@hanmail.net

왔다[1]. 에이전트는 선택할 수 있는 행동의 집합을 가지고 있으며 또한 환경에 대한 많은 변수들이 존재한다. 지능형 에이전트의 학습에 있어서 정해진 전략을 사용하는 교사 학습보다는 환경을 감지하여 스스로 최적의 전략을 세울 수 있는 강화학습과 같은 비 교사 학습이 더욱 효과적이다[2].

본 논문에서는 다수의 자율적인 에이전트들이 팀을 이루어 실시간으로 축구시합을 진행하는 로봇 축구 시뮬레이션 환경에서 퍼지 기반 강화학습 모델을 사용하여 에이전트가 상황에 맞는 적절한 행동을 선택할 수 있도록 최적의 전략을 구성하는 연구를 수행하였다. 로봇 축구 시스템에서는 지능형 로봇들이 효율적으로 협동하여 경기에 이길 수 있도록 적절한 전략을 프로그램하는 것이 중요하다. 로봇 축구 경기의 경우 다양한 상황을 고려한 모델링이 어렵기 때문에 사람이 직접 전략을 프로그래밍 하는 경우 경기에 대한 많은 경험이 필요하게 된다. 이러한 문제를 해결하기 위하여 에이전트들의 빠른 학습 수렴속도와 효율적인 행동 선택 전략에 대한 연구가 필요하다. 본 연구에서는 멀티에이전트의 효율적인 행동선택 방법과 학습 수렴속도 개선을 위하여 퍼지 추론 기반 멀티에이전트 강화학습 모델을 제안하였다.

본 논문의 구성은 다음과 같다. 2장에서는 지능형 에이전트 프레임워크와 멀티에이전트 환경에 대하여 기술하고 3장에서는 강화학습방법과 학습알고리즘과 정책에 관하여 기술하였다. 4장에서는 본 논문에서 제안한 퍼지추론 기반의 멀티 에이전트 강화학습 모델을 제시하였고 5장에서는 로봇컵 Keepaway 테스트베드를 통한 실험으로 제안된 모델의 성능을 평가 하였다. 6장에서는 결론 및 향후 연구 방향에 관하여 기술하였다.

II. 지능형 멀티 에이전트

에이전트는 센서를 통해 주위 환경을 인지하고 주변 환경 정보를 가지고 원하는 목적을 위한 행위를 할 수 있는 존재이다. 에이전트는 인공지능 분야에서 다양한 방법으로 연구되고 있는 개념으로 사실상 인공지능 연구의 최종 목표가 사람과 유사한 지적 능력을 소유한

에이전트의 개발이라고 할 수 있다. 에이전트는 주어진 특정 목적이나 자신이 만든 목적을 이루기 위해 독립적이고 효과적으로 자체 센서나 작동기를 이용한다[2]. 특정 목적을 가진 어떤 기능을 수행하는데 있어서 합리적으로 처리하며 그 성공률은 성능 척도를 통하여 결정된다. 로봇컵의 등장 이후 로봇 축구 시스템은 로봇의 주행제어, 비전 시스템, 멀티 에이전트 등 여러 분야의 연구 소재로서 활용되어 왔으며 에이전트는 체스를 두는 프로그램으로부터 인간이 가기 힘든 미지의 인공 탐사선까지 매우 다양한 분야에 이용되어 진다.

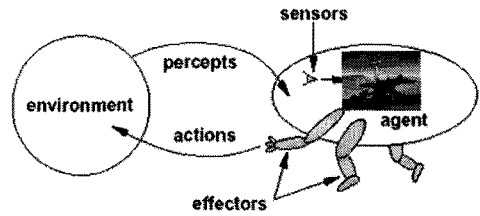


그림 1. 에이전트의 구조

총체적인 지능형 에이전트의 모델로 제시되어 있는 프레임워크로는 Minsky[3]와 Sloman[4]의 모델이 대표적이다. 모든 문제를 해결 할 수 있는 단일한 방법은 없으며 서로 다른 종류의 문제에 각각 다른 방법을 적용하는 것이 필요하다. [그림 2]와 [그림 3]에서 보는 바와 같이 반사적 기능에서부터 학습기능, 정서 기능, 반성적 기능까지 고수준의 지능적 처리를 전제로 한 다양한 기능의 모델을 제시하고 있다.

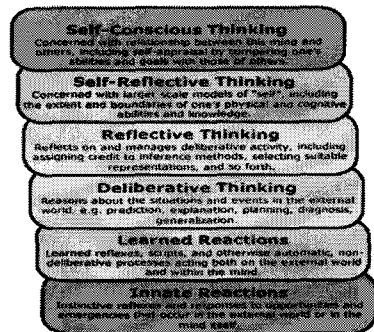


그림 2. Minsky의 지능형 에이전트 프레임워크

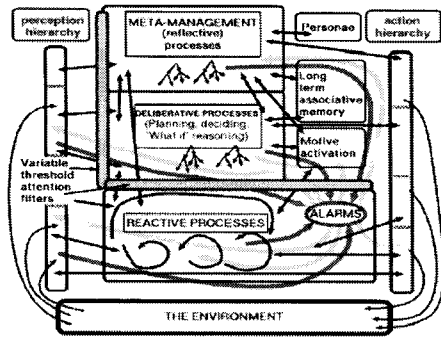


그림 3. Sloman의 지능형 에이전트 프레임워크

멀티 에이전트 시스템은 다수의 에이전트와 이들 사이의 조정을 포함하는 상호작용들이 존재하는 복잡한 시스템이다. 에이전트의 자율적인 행동은 환경과의 상호 작용으로부터 나오며 자율적인 에이전트는 실제 환경에 단독으로 존재 할 수도 있고 다른 에이전트들과 함께 있을 수도 있다. 멀티 에이전트 환경은 하나의 에이전트가 해결하지 못하는 어려운 문제를 해결하기 위해 여러 에이전트들로 구성되는 환경이다[5]. 멀티 에이전트 분야에서의 주요한 쟁점은 상호 협력과 경쟁을 통해 작업을 수행함으로써 최종의 목표를 달성하는 것이다. 경쟁적 관계의 에이전트들의 작업이라면 각각의 에이전트는 서로의 이익만을 위해 최선의 정책을 수행한다. 각각의 에이전트는 특정 도메인에 한정적인 작업을 수행하는 단순 에이전트의 형태를 가지기도 하며 복잡한 작업이 요구되는 경우 다수의 에이전트들이 상호작용과 협력을 통해 작업을 수행한다. 이러한 협동과 경쟁의 문제를 해결하기 위해서는 학습이 필요하며 학습은 단순히 사물을 인식하는 것이 아닌 새로운 정보를 통하여 얻어진 지식을 수정 할 수 있어야 하고 정확한 지식으로 개선해 나가는 능력을 갖도록 하는 것이다. 일반적으로 멀티 에이전트 학습에서는 교사 학습처럼 예제가 주어지지 않고 성취해야 할 목표와 행동을 평가하는 보상 함수가 주어진다.

III. 강화 학습

강화 학습은 에이전트가 수행한 행동에 의해 변화된

환경의 상태에 대해 보상이나 벌칙을 줌으로 하여 스스로 최적의 상태를 학습하도록 하는 방법이다[6]. 에이전트의 학습 목표인 최적의 정책은 상태와 행동의 쌍으로 된 형태이다. 강화학습은 환경과 상호작용하는 과정을 통하여 목표를 이루기 위한 전략을 학습한 방법으로 변화하는 환경에서도 학습이 가능하다는 장점을 가지고 있지만 환경이 커질수록 학습량이 많아져 속도가 느려진다는 단점이 존재한다[7].

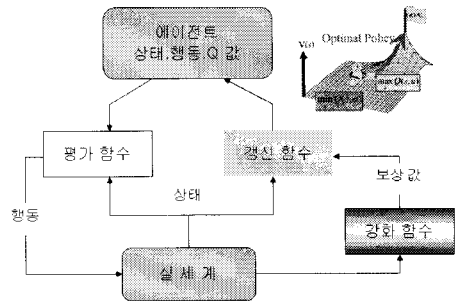


그림 4. 강화학습 Q-Learning 방법

에이전트와 환경의 상호작용은 게임에서의 경기, 미로의 탐험, 반복되는 상호작용과 같은 에피소드들이 있으며 각 에피소드는 정해진 최종 상태에서 끝나게 된다.

Q학습(Q-Learning)은 강화학습 방법 중 가장 널리 이용되는 대표적인 학습 방법으로 Watkins[8]에 의해 제안되었다. Q학습은 상태-행동 쌍에 대한 기대응답의 평가함수를 구하는데 사용된다. 그리고 상태-행동에 대한 평가함수 값을 구하는 방법은 동적 프로그래밍 방식을 사용한다. Q-learning과 같은 비 모델 기반의 강화 학습은 사전에 환경에 대한 별다른 모델을 설정하거나 학습할 필요가 없으며 다양한 상태와 행동들을 충분히 자주 경험할 수만 있으면 최적의 행동전략에 도달할 수 있어 다양한 분야에 적용되고 있다. Q학습의 가장 큰 특징은 환경모델, 즉 보상함수 및 상태 결정함수가 필요 없는 강화학습이라는 점이다. Q학습에서는 상태-행동에 대한 평가함수 $Q(s, a)$ 값을 예측하고 이 Q값을 기초로 행동을 결정한다. 그러나 사전에 정확한 평가함수 $Q(s, a)$ 값을 알 수 없기 때문에 에이전트는 경험을 통해 점진적으로 최적의 Q함수 값을 찾아간다. 최적의

Q함수 값에 수렴하기 위해서는 가능한 모든 상태와 행동에 대한 충분한 반복경험이 필요하다.

IV. 퍼지 추론기반 멀티에이전트 강화학습

퍼지 추론 시스템(Fuzzy Inference System)은 상태 변수의 조건에 따라 적합한 행동을 결정하는 규칙의 집합이다. 퍼지 추론 시스템은 보편적으로 사용되고 있는 함수 근사 방법이고 Q-함수를 저장하기에 적합하다. 일반적으로 If ~Then~의 형태로 구성되어 사람이 이해하기가 용이하며 또한 사전 지식을 퍼지 규칙에 포함시킬 수 있어 학습 시간을 줄일 수 있는 장점이 있다[9].

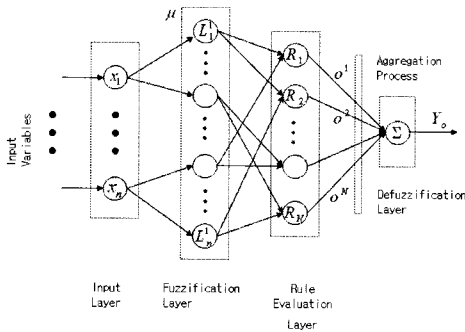


그림 5. 퍼지 추론 시스템의 구조

퍼지 추론 시스템은 [그림 5]에서 보는 바와 같이 첫째 층은 입력 값들을 받아들인다. 둘째 층에서 입력 값의 퍼지 레이블에 대한 소속도가 구해진다. 세 번째 층에서 규칙의 진리값을 구하고 규칙들의 진리값과 결론부를 바탕으로 출력이 계산되어 진다.

퍼지 추론 시스템을 정의할 때 사용자는 입력 변수를 결정하고 퍼지 집합의 형태와 개수를 정해야 한다. 입력 상태 변수는 연속적인 값을 퍼지 멤버 함수(Fuzzy membership function)를 이용하여 이산적인 언어 변수의 형태로 표현한다. 퍼지 소속 함수는 다양한 형태가 있으며 삼각 또는 사다리꼴 형태의 퍼지 멤버 함수는 계산이 간단하지만 미분이 가능하지 않는 단점이 있고 가우시안 또는 시그모이드 함수는 미분이 가능하나 계산량이 많다[10].

퍼지 강화 학습은 퍼지 추론 시스템을 강화 학습을 통해 해당 시스템에 적합하도록 학습 시키는 알고리즘이다. 퍼지 강화 학습은 강화 학습이 이산적인 상태 변수만을 인식 할 수 있고 이산적인 출력 변수만을 나타내는 단점을 보완하기 위해 제안되었다[11]. 퍼지 강화 학습은 강화 학습의 특징과 퍼지 추론 시스템의 특징을 모두 가지고 있어서 Q-learning과 같이 정확히 모델링 할 수 없는 환경에 대해 동적인 학습이 가능한 장점이 있다. 또한 퍼지 추론 시스템을 사용하여 사람이 사전 지식을 미리 입력 할 수도 있으며 연속적인 변수들을 사용할 수 있다. [그림 6]에서 퍼지 추론을 이용한 자동온도조절 시스템에서 의사결정 과정의 예를 보여주고 있다.

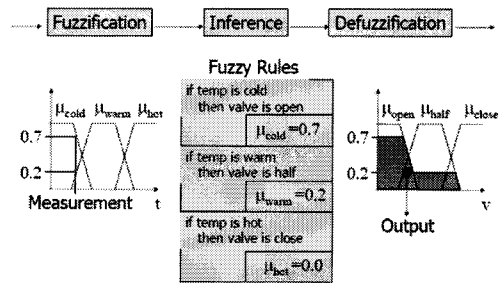


그림 6. 자동온도조절 시스템에서의 퍼지추론 과정

퍼지 추론 시스템에서 규칙의 입력 상태 변수와 출력 변수의 정의는 사용자가 미리 해주게 된다. 그러나 주어진 시스템에 대해 최적의 규칙을 정의하기 어려울 경우 정의된 모든 규칙들의 가능한 출력 변수를 여러 개로 정의하여 놓고 이 중에서 최적의 출력 변수를 학습을 통하여 찾도록 한다. 본 논문에서는 에이전트의 효율적인 행동 선택 전략을 수행 위하여 에이전트 상호간의 거리와 시야를 기준으로 소속도를 구하고 퍼지규칙을 생성하여 행동 선택을 결정하였다. [그림 7]은 축구 에이전트의 퍼지 추론 메커니즘을 보여주고 있다.

본 논문에서는 퍼지 추론 시스템과 강화학습 알고리즘을 기반으로 로봇축구 경기에서 멀티 에이전트의 효과적인 행동 선택과 학습의 수렴속도를 개선하기 위하여 퍼지 추론 시스템과 강화학습을 이용한 퍼지추론 기반의 멀티에이전트 강화학습 모델을 제안하였다.

keeper와 taker의 개수이다.

Keepaway는 에피소드 방식으로 이루어지며 keeper가 공을 빼앗기거나 영역 밖으로 나갔을 때가 하나의 에피소드가 된다. 각각의 에이전트는 공과 상대편 에이전트 등의 정보를 받아들이며 가장 기본적인 행동인 kick, turn, dash를 수행한다.

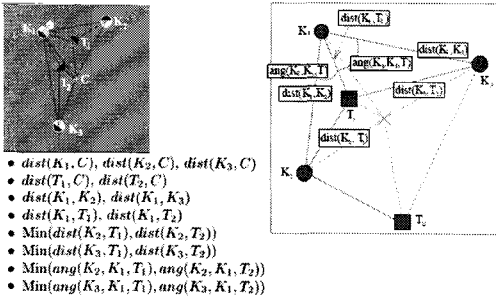


그림 10. keepaway 상태 변수

3대2 Keepaway의 경우 [그림 10]에서 보는 바와 같이 13개의 상태 변수를 가진다. Keepaway에서 각각의 에이전트는 로보컵 시뮬레이션 리그의 기본적인 행동과 [표 1]에서 보이는 바와 같이 상위 레벨의 5가지의 매크로 행동을 선택한다. 행동 선택은 하나의 상태에서 각 keeper가 취할 행동들에 대한 Q값들을 계산하고 그 값들을 기준으로 비율을 정하며 이를 바탕으로 행동이 선택될 수 있도록 하였다. 에피소드를 단위로 에피소드의 기간이 얼마나 지속되는가를 학습의 측정치로 사용하였다.

표 1. CMUnited-99 Keepaway 상위 매크로 행동

Macro Action	내용
HoldBall()	가능한 상대편에게 멀리 떨어져서 공을 점유하고 유지하는 동안 움직이지 않는 행동
PassBall(k)	keeper k 앞으로 공을 직접 치는 행동
GoToBall()	공을 가로채기하거나 정지된 공 앞으로 이동
GetOpen()	공이나 상대편으로부터 자유로운 위치로 이동
BlockPass(k)	공을 가지고 있는 keeper와 Keeper 사이로 이동

[그림 11]은 Keepaway 환경에서 강화학습을 통한 에이전트를 학습을 나타낸 것이다. 그림에서 보는 바와 같이 학습시간이 지속됨에 따라 학습의 효율성이 향상됨을 확인할 수 있다.

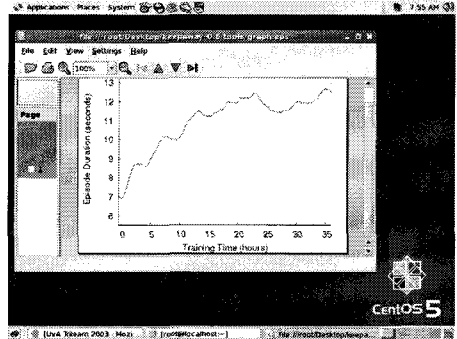
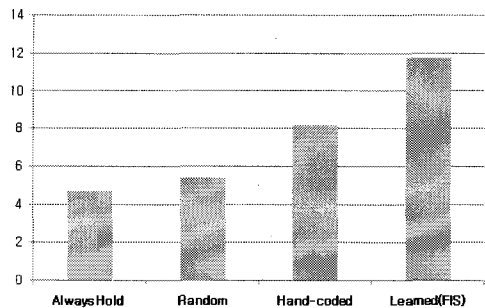


그림 11. keepaway Learning Graph

본 논문에서 제안한 강화학습 모델의 학습 행동 정책의 효율성 평가를 위하여 Keeper의 정책을 Always Hold, Random, Hand-coded, Learned(FIS)로 구분하여 비교 실험하였다. 각각의 규칙은 다음과 같다.

- Always Hold : 항상 HoldBall() 수행
- Random : HoldBall() or PassBall(k) 수행
- Hand-coded : Taker가 4m 안에 없으면 HoldBall()을 수행하고 Teammate에게 볼을 성공적으로 패스할 수 있는 위치에 있을 경우 PassBall(k) 수행 (CMUnited-99 pass-evaluation function 사용)
- Learned(FIS) : 퍼지 추론 메커니즘에 의해 아군의 거리와 시야에 따라 소속도를 구하고 퍼지 규칙을 생성하여 행동선택 수행



Keeper Policy	Average Ball Possession time
Always Hold	4.7(±0.2)Sec
Random	5.4(±0.3)Sec
Hand-coded	8.2(±0.6)Sec
Learned(FIS)	11.8(±0.5)Sec

그림 12. 학습 정책 별 Keeper의 평균 볼 소유시간

행동 정책의 효율성을 평가는 Keeper의 평균 볼 점유 시간으로 측정하였다. 총 50시간씩 10번에 걸쳐 실험하였으며 에피소드는 약 100만개이다.

Keeper의 행동정책 비교실험 결과에서 [그림 12]에 서와 같이 Keeper의 평균 볼 점유시간이 제안한 퍼지 추론 기반의 강화학습모델을 적용했을 경우 다른 세 가지 정책을 선택하였을 경우에 비해 Keeper의 볼 소유 시간이 향상됨을 보였다. 따라서 강화학습에 있어 퍼지 추론을 이용한 사전지식을 적용함으로써 에이전트의 행동선택의 효율성을 높일 수 있음을 확인하였다. 본 논문에서 제안한 퍼지 추론 기반의 강화학습 모델은 실험에서와 같이 제한된 에피소드 방식의 멀티 에이전트 학습 환경에서 효율적인 행동 선택을 보였으나 환경에 대한 지식이 예측하기 힘들고 불확실하며 연속적인 상태공간이 지속되는 환경에서는 제한된 모델의 적용이 어려우며 새로운 학습정책과 방법의 적용이 요구된다.

VI. 결론

에이전트는 사용자의 목적과 주위 환경으로부터 최적의 행동을 스스로 찾아낼 수 있어야 한다. 본 논문에서는 로봇 축구경기에서 에이전트의 효율적인 경기 수행을 위해 퍼지기반의 멀티에이전트 강화학습 모델을 제안하였다. 에피소드 방식의 강화학습 환경인 로보컵 Keepaway를 테스트베드로 삼아 제안한 퍼지 추론 기반 강화학습 모델을 적용하여 에이전트가 로봇 축구 경기에서 상황에 맞는 적절한 행동 선택과 빠른 학습 수렴속도를 갖도록 하였다. 제안된 퍼지 추론 기반의 멀티에이전트 강화학습모델은 지능형 멀티 에이전트의 학습에서 행동 정책의 효율성 평가와 로봇축구 시스템의 전략 및 전술에 활용이 가능하다. 향후 연구로는 논문에서 사용된 제한된 에피소드 환경이 아닌 온라인 게임 환경과 같이 연속적이며 지속적인 점진적 방식의 환경에서 에이전트의 효율적인 행동정책에 관한 연구를 수행 하고자 한다.

참고 문헌

- [1] E. Matthew, Taylor and Peter Stone, "Representation Transfer for Reinforcement Learning," AAAI 2007 Fall Symposium, Arlington, Virginia, 2007.
- [2] E. Yang, and Gu., "Multi-Agent Reinforcement Learning for Multi-Robot System: A Survey," University of Essex Technical Report CSM-404, 2004.
- [3] Marvin Minsky, Push Singh, and Aaron Sloman, "The St. Thomas Common Sense Symposium: Designing Architectures for Human-Level Intelligence," AI Magazine, 25-2: Summer 2004, pp.113-124. 2004.
- [4] Aaron Sloman and Ron Chrisley, "Virtual Machines and Consciousness," Journal of Consciousness Studies 10, 4-5. 2003.
- [5] 권기덕, 김인철, "적대적 멀티 에이전트 환경에서 효율적인 강화 학습을 위한 정책 모델링", 한국정보과학회 논문지: 소프트웨어 및 응용 제35권, 제 3호, pp179-188, 2008.
- [6] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," MIT Press, 1998.
- [7] G. Tesauro, "Multi Agent Learning: Mini Tutorial," IBM Watson Research Center, 2000.
- [8] C. J. C. H. Watkins and P. Dayan, "Technical notes: Q-learning," Machine Learning, Vol.8, pp.279-292, 1992.
- [9] R. Fagin, "Combining Fuzzy Information from Multiple Systems," J. of Computer and System Sciences, Vol.58, pp.83-99. 1999.
- [10] 이영아, 정경숙, 정태충, "퍼지 클러스터링을 이용한 강화학습의 함수근사", 한국정보처리학회 논문지, 제10-B권, 제6호, pp.587-592, 2003.
- [11] Lionel Jouffe, "Fuzzy Inference System Learning by Reinforcement Methods," IEEE Transactions on Systems, Man and Cybernetics pp.338-355,

1998.

[12] David McAllester and Peter Stone. Keeping the ball from CMUnited-99, Robocup-2000:Robocup IV, Springer Verlag, Berlin, 2001.

[13] Alexander A. Sherstov and Peter Stone, Function Approximation via Tile Coding: Automating Parameter Choice, In Proc. Symposium on Abstraction, Reinforcement, and Approximation(SARA-05), 2005.

스토리지 보안, 데이터베이스 보안, 사이버공격 및 침입탐지

저자 소개

이 봉근(Bong Keun Lee)

정회원



- 1997년 : 한남대학교 컴퓨터공학과(공학사)
- 1999년 : 한남대학교 컴퓨터공학과(공학석사)
- 2000년 ~ 현재 : 충북대학교 전기전자 컴퓨터공학부 박사과정

• 2006년 ~ 현재 : (주)에이티엔 기술연구소 연구원
 <관심분야> : 지능형 에이전트, 상황인지, 게임인공지능, 데이터마이닝, 객체 및 지식기반 시스템, 데이터베이스 보안, 유비쿼터스 컴퓨팅

정재두(Jae Du Chung)

정회원



- 1987년 : 연세대학교 전산학과(이학사)
- 1991년 : 미해군대학원 전산전공(공학석사)
- 2004년 : 충북대학교 대학원 전산전공(공학박사)

• 1982년 ~ 1984년 : 한국국방연구원 연구원
 • 1991년 ~ 2001년 : 육군본부중앙전산소 운영팀장
 • 2002년 ~ 2006년 : 육군본부중앙전산소 전산소장
 • 2007년 ~ 현재 : 국방부 전산장비 운영센터장
 <관심분야> : 공간 데이터베이스, Temporal GIS, 통합형 전산자료구축, 지식기반 정보검색 시스템, 서버 및

류근호(Keun Ho Ryu)

정회원



- 1976년 : 숭실대학교 전산학과(이학사)
- 1980년 : 연세대학교 대학원 전산전공(공학석사)
- 1998년 : 연세대학교 대학원 전산전공(공학박사)

- 1976년 ~ 1986년 : 육군군수 지원사 전산실(ROTC장교), 한국전자통신연구원(연구원), 한국방송통신대 전산학과(조교수) 근무
- 1989년 ~ 1991년 : Univ. of Arizona Research Staff (TempIS 연구원, Temporal DB)
- 1986년 ~ 현재 : 충북대학교 전기전자 컴퓨터공학부 교수

<관심분야> : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 객체 및 지식기반 시스템, 지식기반 정보검색시스템, 데이터마이닝, 데이터베이스 보안 및 Bio-Informatics