Proceedings of KFIS 2001 Spring Conference, 2001. 5.

인공생명의 연구에 있어서 강화학습의 전략

Strategy of Reinforcement Learning in Artificail Life

심귀보 · 박창현
중앙대학교 전자전기공학부

Kwee-Bo Sim and Chang-Hyun Park
School of Electrical and Electronic Engineering, Chung-Ang University
E-mail : kbsim@cau.ac.kr

ABSTRACT

일반적으로 기계학습은 교사신호의 유무에 따라 교사학습과 비교사학습, 그리고 간접교사에 의한 강화학습으로 분류할 수 있다. 강화학습이란 용어는 원래 실험 심리학에서 동물의 학습방법 연구에서 비롯되었으나, 최근에는 공학 특히 인공생명분야에서 뉴럴 네트워크의 학습 알고리즘으로 많은 관심을 끌고 있다. 강화학습은 제어기 또는 에이전트의 행동에 대한 보상을 최대화하는 상태-행동 규칙」이나 행동발생 전략을 찾아내는 것이다. 본 논문에서는 최근 많이 연구되고 있는 강화학습의 방법과 연구동향을 소개하고, 특히 인공생명 연구에 있어서 강하학습의 중요성을 역설한다.

Keywords : Reinforcement Learning, Q-learning, Temporal-Difference Learning, Artificial Life

Ⅰ. 서 론

일반적으로 기계학습은 교사신호의 유무에 따라 교사학습과 비교사학습, 그리고 간접교사에 의한 강화학습으로 분류할 수 있다. 강화학습이란 용어는 원래 실험 심리학에서 동물의 학습방법 연구에서 비롯되었으나, 최근에는 공학 특히 인공생명 분야에서 뉴럴 네트워크의 학습 알고리즘으로 많은 관심을 끌고 있다. 강화학습은 일반적으로 제어기 또는 에이전트의 행동에 대한 보상(reward)을 최대화하는 「상태-행동(state-action) 규칙」이나 「행동발생 전략」을 찾아내는 것이다. 그러나 많은 실세계의 경우에 있어서 목표에 도달할 때까지는 중간 단계의 행동에 대한 즉각적인 보상이 주어지지 않는다. 이러한 경우 외부로부터의 강화신호가 없기 때문에 학습이 일어나지 않게 된다. 그러한 경우에도 목표에 도달하기 위해서는 지속적인 학습이 이루어져야 하므로 일시적인 신뢰할당이 이루어져야 한다. 이것을 「신뢰할당 문제(credit-assignment problem)」라고 하며, 강화학습에 있어서 가장 중요한 문제라고 할 수 있다. 이 문제에 대한 가장 일반적인 해결 방법은 강화 신호를 생성하는 외부 평가함수보다 더 자세한 정보를 얻을 수 있는 내부 평가함수를 구현하는 것이다.

본 논문에서는 이러한 강화학습의 개념과 연구 방법들을 소개하고, 특히 인공생명 분야의 연구에 있어서 강하학습의 중요성과 전략에 대해서 기술한다.

Ⅱ. 강화학습의 개념

먼저 그림 1과 같이 미지의 환경에 놓인 로봇과 같은 에이전트를 생각하자. 이 에이전트는 환경으로부터의 센서 입력에 대해서 자신의 행동을 선택하여 실행한다. 일련의 행동에 대해서 환경으로부터 보상이 주어진다. 여기서 보상이란 에이전트의 존재 의의를 추상화한 량이라고 할 수 있다. 예를 들어, 생물인 경우는 먹이를 취하거나 적으로부터 달아날 때, 로봇인 경우는 목표를 달성할 때에 주어지는 양이다. 일반적으로 에이전트는 환경의 모든 상태변수를 감지할 수 있는 것이 아니기 때문에, 불확실성의 처리가 필수적이다. 강화학습에서 주어지는 보상은 행동 하나 하나의 좋고 나쁨을 교시하는 것이 아니라, 일련의 행동의 결과에 대해서 주어지기 때문에 지연 처리가 요구된다. 이와 같이 강화학습에서는 「불확실성」과 「지연보상」을 갖는 미흡한 정보원밖에 이용할 수 없다는 특징이 있다.

보통 에이전트는 그림 2와 같이 상태 인식부, 행동 선택부 그리고 학습부 등 3개의 구성요소로 되어 있다. 상태 인식부는 센서 입력으로부터 행동의 후보가 되는 경합 집합을 만든다. 예를 들어, 규칙 베이스인 경우 규칙의 조합이 이에 해당된다. 행동 선택부는 경합 집합으로부터 행동을 선택한다. 예를 들어, 규칙의 하중에 비례하는 확률로 행동을 선택한다. 학습부는 보상에 따라서 상태 인식부가 보다 적절한 행동의 후보를 생성하도록 갱신한다.

강화학습은 교사가 없는 학습법의 일종이다. 그림 2에 교사학습과의 비교를 나타낸다. 전문가 시스템(expert system)의 지식 획득과 같은 교사학습에서는, 에이전트는 동일한 목적과 올바른 정답을 가지고 있는 교사와 정보를 교환한다. 교사가 있는 한은 확실성과 지연이 없는 평가에 의해 서로간의 협력을 기대할 수 있다. 반면에 강화학습에서는, 에이전트는 보상의 극대화라고 하는 목적만을 가지고 환경과 정보를 교환한다. 이 강화학습에서는 환경으로부터의 협력은 기대할 수 없기 때문에 불확실성과 지연보상은 본질적으로 존재하고, 충분한 정보를 획득할 때까지는 시행착오를 반복하지 않으면 안 된다. 하지만 서투른 교사에 의한 것보다는 우수한 기능을 획득할 가능성이 있다. 특히 강화학습은 최근의 인공생명 연구의 중요한 주제인 협조현상의 창발이라고 하는 애초에 정답이 없는 문제의 접근에 길을 열고 있다.

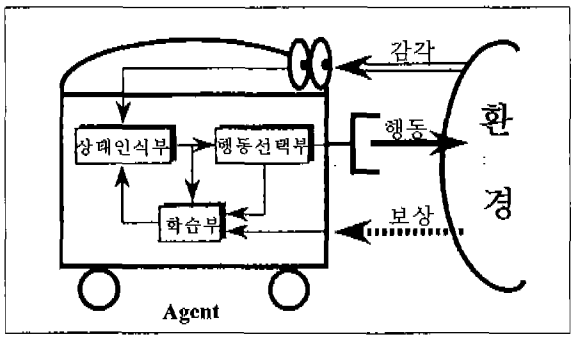

그림 1. 강화학습의 패러다임

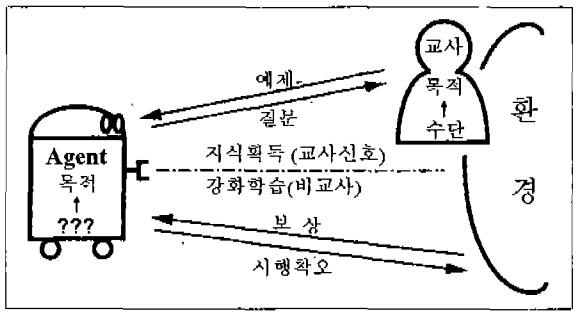

그림 2. 교사학습과 강화학습

III. 강화학습의 분류

강화학습의 연구를 분류하여 앞으로의 발전 방향을 모색하기 위해서 일종의 맵을 생각해 보자. 분류의 한 축은 「환경의 클래스」이고, 다른 한 축은 「접근의 지향성」이다. 환경의 클래스는 상태천이에 마르코프성(Markov)을 가정하는가 하지 않는가에 따라 나누어진다. 그 이유는 마르코프 과정에는 많은 연구가 축적되어 있어서 강력한 해석 수단을 제공하고 있기 때문이다. 그리고 강화학습에서 요구되는 성능은, 결과로서 가능한 한 많은 보수를 얻어야 하는 「최적성」과, 학습 도중에도 가능한 한 보수를 계속 얻어야 하는 「효율성」의 2가지 측면이 있다. 최적성은 환경에 대해서 가능한 한 넓은 범위를 탐색함으로서 얻을 수 있기 때문에, 최적성을 중시하는 접근을 환경 동정형(exploration oriented)이라 부른다. 한편 효율성은 보상을 얻은 경험을 분석하고 반복함으로써 얻을 수 있기 때문에, 효율성을 중시하는 접근을 경험 강화형(exploitation oriented)이라 부른다. 기존의 대표적인 강화학습의 연구를 이들 두 개의 축에 따라서 위치시켜보면 그림 3과 같다. 그림에서 화살표는 개선, 발전의 방향을 나타낸 것이다.

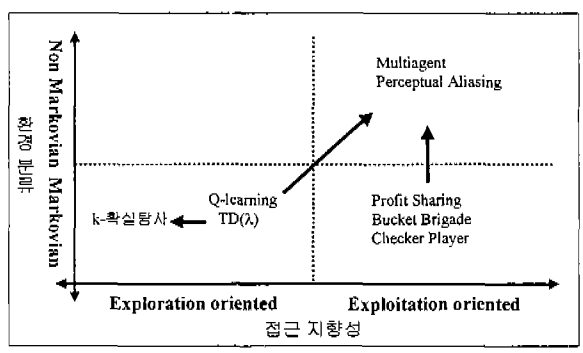

그림 3. 강화학습의 분류

3.1 환경동정형 학습

환경동정형 학습에서는 그 성격상 동정해야 하는 환경의 클래스를 가정하지 않으면 안 된다. 불확실성과 지연보상이 없는 클래스를 가정할 수도 있지만, 이들 클래스의 동정 문제는 이미 확립되어 있다. 마르코프 결정과정은 불확실성과 지연보상을 포함하는 가장 단순한 클래스이고, 1960년대부터 활발하게 연구되고 있다. 마르코프 결정과정의 환경은 그림 4와 같은 상태 천이도로 표현된다. 그림에서 노드는 상태로써 서로 다른 감각 입력에 해당한다. 화살표로 표현된 가지는 상태 천이로써 선택 가능한 행동에 해당한다. 가지가 나누어지는 것은 상태천이의 불확실성을 나타내고 삼각형으로 표시된 위치에서 보상이 주어진다.

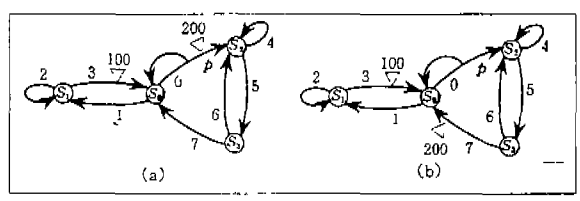

그림 4. Markovian 환경의 예

대표적인 환경동정형 학습법에는 그림 3에서

알 수 있는 바와 같이 Temporal-Difference법(이하 TD(λ)법이라 한다)과 Q-learning, 그리고 k-확실탐사 등이 있다. TD(λ)법은 각 상태의 평가를 동정하는 방법인데 λ는 decay-rate parameter(for eligibility traces)로 0~1까지의 값을 갖는다. 일반적으로 λ가 0이면 TD(0)인 고전적 dynamic programming에 가깝고, TD(1)은 교사학습에 가깝다.

Q-learning은 TD(λ)법의 발전형으로서 상태뿐만이 아니라 상태-행동을 나타내는 짝(pair)의 평가를 동정한다. Q-learning의 구성을 그림 1에 적용시켜 설명하면, 상태 인식부는 상태-행동 조로 구성된 규칙 테이블(룰 베이스)이고 각 규칙은 Q값을 가지고 있다. 행동 선택부에는 Q값에 기초한 룰렛 선택이나, 행동의 90%는 최대의 Q값을 가지는 규칙을 선택하고 나머지 10%는 랜덤하게 선택하는 등의 다양한 것이 이용된다. 그리고 학습부에서는 다음 식에 따라 Q값을 갱신한다.

$$Q_{t+1}(s,a) = (1-\alpha_t)Q_t(s,a) + \alpha_t(r + \gamma \max_{a' \in A} Q_t(s',a'))$$

어떤 스케쥴에 따라서 학습율 α를 감소시키면서 여러 번 시행 후에 Q값이 수렴하면 각 상태에서 최대의 Q값을 갖는 규칙을 선택하면 그것이 최적의 정책(policy)이 된다. Q-learning의 이점은 환경이 마르코프적이라면 최종적으로 최적인 행동결정 전략이 얻어질 수 있다는 것이다. 반면 결점으로는 해석이 보증하고 있는 것은 어디까지나 최종 결과인 점과 해석이 에이전트의 세 개의 구성 요소 중에 행동 선택부를 포함하고 있지 않은 점이다. 그 결과 경우에 따라서는 상당히 낭비가 되는 시행을 수반하여 학습에 시간이 걸리고, 학습 도중에 Q값이 환경의 구조나 학습률 등의 파라미터에 민감할 뿐 아니라 동적으로 변화하는 환경에서 불안정한 거동을 나타낸다. 그렇지만 Q-learning은 비교적 간단하기 때문에 자율이동로봇을 중심으로 많이 이용되고 있는 학습법이다. 그런데 실제로 응용하기 위해서는 연속량의 취급이 매우 중요하다. 따라서 최근에는 연속의 감각입력을 취급하기 위해서 상태 인식부를 뉴럴 네트워크 등을 이용해서 Q값을 함수 근사하는 연구도 행해지고 있다.

한편 Q-learning의 발전형은 이들의 결점의 극복하는 것을 목표로 하고 있다. 그 중의 하나에 k-확실 탐사법이 있는데 이는 극단적인 환경동정을 지향하는 학습법이다. k-확실 탐사법에 기초한 학습시스템을 그림 1에 적용하여 설명하면, 상태 인식부는 Q-learning과 동일한 룰 베이스이다. 그러나 Q값 대신에 어떤 상태인 행동을 취했을 때 다른 상태로 천이하는 상태천이 확률의 추정값과 보수의 기대치인 통계정보를 가지고 있다. 이 통계 정보로부터 policy iteration에 의해 정책을 구할 수 있는데, 정책의 최적성은 현재의 통계 정보가 어느 정도 정확한가에 의존한다.

통계적 추정의 정확성은 시행 회수에 의해서 단조 증가하기 때문에 추정의 정확성을 나타내기 위해서 k-확실이라는 개념을 이용한다. 어떤 룰이 k-확실하다고 하는 것은 선택 회수가 k회 이상인 것을 의미한다. 룰 베이스는 이 k를 관리하고 있고, 행동 선택부는 k-확실에 도달하지 않는 룰을 적극적으로 선택함으로써 추정의 정도를 높인다. 학습부는 통계 데이터를 계속 집계하여 모든 룰이 k-확실로 되었을 때 k를 1증가시킨다. 따라서 k가 충분히 크게되어 통계적 추정이 정확하게 될 때 최적의 행동이 보증된다. 그러나 규칙 베이스 자체가 상당히 클 경우에는 1-확실에 도달하는 데도 막대한 시행 회수가 필요하게 된다.

2.2 경험강화형 학습

고전적인 강화학습은 모두 경험강화형이다. 대표적인 것을 세 가지 예를 들어 간단히 설명한다. 첫 번째는 유명한 Samuel의 "Checker player"가 있는데 이것은 게임의 나무의 탐색에 사용하는 평가함수를 게임의 승패를 보수로서 강화한다. 2대의 에이전트가 서로 게임을 하면서 학습하기 때문에 학습하는 멀티 에이전트시스템이라고 한다. 둘째로 "Bucket brigade"이 있는데 이것은 행동을 선택할 때마다 내기(bet)를 한다. 어떤 단계에서 경합하는 룰은 일정한 내기 금액을 지불한다. 룰렛에 의해서 선택된 룰은 승자로 간주되어 보수와 다음 단계의 내기 금액의 합계가 주어진다. 마지막으로 "Profit sharing"이 있는데 이것은 경합에 참가한 룰에 보수를 어떤 일정한 방법으로 분배한다. 이름 그대로 이득 분배를 하는 방법이다.

고전적인 경험강화형 학습은 대개 즉흥적인 아이디어라고 하는 측면을 가지고 있고, 그 거동도 파라미터에 민감하여, 성공, 실패, 진동 등 경우에 따라 달라진다. 원래 경험강화형 학습에서 요구되는 성능은 환경동정형과 같이 exploration에 에너지를 소비하지 않고 계속적으로 보수를 얻는 행동 패턴을 확립하는 것이다. 따라서 경험강화형 학습을 실용화하기 위해서는 일정의 합리성을 보증할 필요가 있다.

한편 강화학습에서 비 마르코프적 환경으로의 지향은 앞으로의 강화학습의 발전에 매우 중요하다. 한 마디로 비 마르코프적 환경이라고 해도 다양한 종류가 있다. 여기서는 그 중에 두 가지의 예를 들어 본다.

어떤 에이전트의 환경이 마르코프적으로 상태천이를 하고 있다고 하더라도 그 에이전트의 감각(센싱) 기능의 제약 때문에 서로 다른 상태

를 구별할 수 없는 경우 에이전트에 있어서는 환경이 비 마르코프적으로 보일 때가 있다. 이것을 "perceptual aliasing"이 있다고 한다. 그림 5는 perceptual aliasing이 있는 환경의 예를 나타낸다. 에이전트는 인접 8개의 격자밖에 인식할 수 없기 때문에 1a와 1b 등의 위치는 에이전트가 구별할 수 없다. 특히 1a와 1b는 최적 행동이 각각 상·하 역으로 되어 있기 때문에 Q-learning을 그대로 적용한다면 동작은 불안정하다. 실제로 많은 경우에 있어서 이와 같은 문제가 발생할 수 있다. 이런 경우에는 적절하게 환경 모델을 구축해 가면서 profit sharing에 의해서 강화학습을 하거나 시계열 데이터 처리에 우수한 리커런트 뉴럴 네트워크(recurrent neural network) 등을 사용한 강화학습 등이 연구되고 있다. 또 하나는 멀티 에이전트 시스템을 들 수 있다. 이 예는 복수의 포식자가 협력해서 피식자를 잡아먹는 문제인데 이것은 동물 집단의 협조 행동의 전형적인 한 예이다. 포식자의 학습모델로서 Q-learning을 사용할 경우, 일반적으로 Q-learning은 동적인 환경에서는 불안정한 거동을 보이기 때문에 통신 등의 보조적 수단이 필요하게 된다.

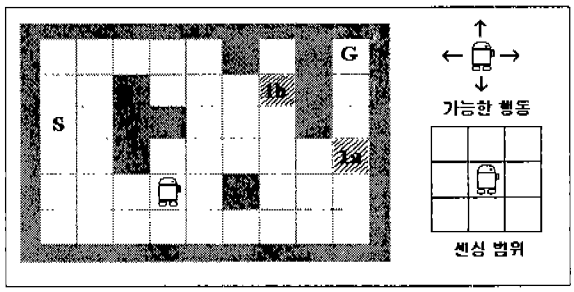

그림 5. Perceptual aliasing이 있는 환경

Ⅳ. 결 론

적응행동을 실현하기 위해서 실제의 동물이 이용하고 있는 기능의 하나가 학습이다. 인공지능 분야에서도 재귀 학습, 연역학습, 개념형성 등, 학습의 연구가 활발하지만 동물학습의 모델로서는 강화학습의 구조가 가장 적합한 것으로 생각된다. 학습자는 보상과 벌칙으로부터 그 원인이 된다고 생각되는 자신의 행동에 대해서 얼마간의 평가를 줌으로서, 보상은 많이 벌칙은 적게되도록 행동규범을 변경한다. 본래는 강화신호인 보상과 벌칙은 학습자의 외부로부터, 예를 들면 환경에 의해 주어진 것이라고 하는 의미가 있지만, 인공생명의 입장에서는 오히려 입력에 관한 학습자 측의 해석으로서 좋고 나쁨을 판단하는 것으로 생각하는 편이 자연스럽다.

일반적으로 동물이나 로봇의 학습에서는 보상이나 벌칙은 어떤 길이의 행동계열의 실행결과로서 얻어지는 경우가 많아서 학습자의 의사결정에 대해서 그 직후에 평가가 얻어지는 경우는 없다. 교사신호에 의한 패턴 분류학습 등 다른 많은 구조에서는 학습자의 의사결정의 직후에 그에 대한 평가, 즉 정답인가 아닌가가 주어지지만, 강화학습에서 다루는 학습문제에서는, 행동에 대한 평가의 시간적인 지연을 수반하는 문제점이 있다. 또한 기계학습의 분야에서 행해지는 강화학습 연구는 이산 시간에서의 입력의 샘플링과 의사결정과 행동의 반복이라고 하는 구조에 기반하는 것이 대부분이다. 이것은 수리적으로는 마르코프 결정문제에 해당한다. 그러나 최근에는 좀더 생명현상에 가까운 비 마르코프 환경에서의 학습에 대한 연구가 활발히 진행되고 있다. 뿐만 아니라 최근에는 그림 7과 같이 진화연산을 이용하여 학습과의 상호보완을 시도한 진화적 강화학습(Evolutionary Reinforcement Learning)도 연구되어 있는데, 이것은 인공생명 연구의 입장에서 학습능력의 진화나 학습능력이 진화의 과정에 미치는 효과에 대한 연구이다.

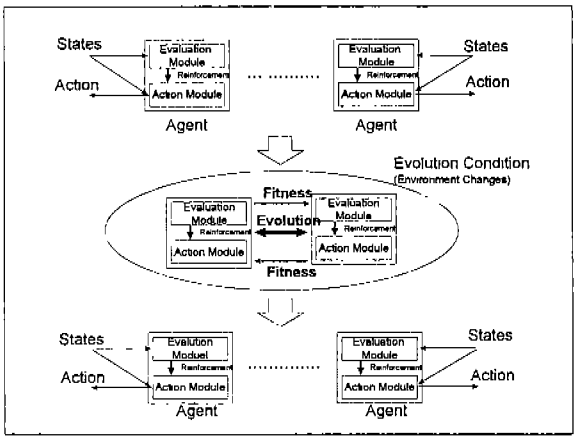

그림 6. 진화적 강화학습

감사의 글 : 본 연구는 과학기술부의 뇌과학 프로젝트(Braintech 21)의 지원에 의하여 이루어진 결과임.

참고문헌

[1] Richard S. Sutton and Andrew G. Barto, *Reingorcement Learning*, The MIT Press, 1998.

[2] 전호병, 심귀보,"학습에 의한 진화전략의 수렴성에 관한 연구", *한국퍼지 및 지능시스템학회 논문지*, vol. 9. no. 6 pp.560-656, 1999.

[3] 박영철, 전효병, 심귀보, "강화학습과 조건적 진화에 의한 자율이동로봇군의 협조행동", *제14차 한국자동제어학술회의(KACC '99) 논문집*, pp. B-144-147, October 14. 1999.