지능형 에이전트의 환경 적응성 및 확장성에 대한 연구

백혜정, 박영택
숭실대학교 컴퓨터 학부
baekhj@korea.com, park@computing.ssu.ac.kr

The study on environmental adaptation and expansion of the intelligent agent

Baek Hae-Jung, Park Young-Tack
Dept. of Computing, Soongsil University.

요 약

로봇이나 가상 캐릭터와 같은 지능형 에이전트가 자율적으로 살아가기 위해서는 주어진 환경을 인식하고, 그에 맞는 최적의 행동을 선택하는 능력을 가지고 있어야 한다. 본 논문은 이러한 지능형 에이전트를 구현하기 위하여, 외부 환경에 적응하면서 최적의 행동을 배우고 선택하는 방법을 연구하였다. 본 논문에서 제안한 방식은 강화 학습을 이용한 행동기반 학습 방법과 기호 학습을 이용한 인지 학습 방법을 통합한 방식으로 다음과 같은 특징을 가진다. 첫째, 외부 환경의 적응성을 수행하기 위하여 강화 학습을 이용하였으며, 이는 지능형 에이전트가 변화하는 환경에 대한 유연성을 가지도록 하였다. 둘째, 경험들에서 귀납적 기계학습과 연관 규칙을 이용하여 규칙을 추출하여 에이전트의 목적에 맞는 환경 요인을 학습함으로 주어진 환경에서 보다 빠르게, 확장된 환경에서 보다 효율적으로 행동을 선택을 하도록 하였다. 제안한 통합 방식은 기존의 강화 학습만을 고려한 학습 알고리즘에 비하여 학습 속도를 향상 시킬수 있으며, 기호 학습만을 고려한 학습 알고리즘에 비하여 환경에 유연성을 가지고 행동을 적용할 수 있는 장점을 가진다.

1. 서 론

오래 전부터, 차세대 사용자 인터페이스로서 로봇이나 가상 캐릭터와 같은 지능형 에이전트에 대한 연구가 진행되어 왔다[1][10]. 이러한 지능형 에이전트는 주어진 환경을 인식하고, 필요한 목적을 성취하기 위하여 적절한 행위를 수행하는 능력을 가진다[5]. 이처럼, 주어진 환경에 능동적으로 대처하는 환경 적응성을 위해서 강화 학습을 이용한 환경 적응에 대한 연구가 많이 이뤄져 왔다. 여기서 강화 학습은 현재 상황에 대한 행동을 수행하고 이에 따라 보상을 받음으로 현재 상황에 가장 적절한 행동을 학습하는 방법으로 환경에 대한 사전 지식 없이 행동을 학습하는 장점을 가진다. 하지만 강화 학습은 최적의 상태와 행동을 찾아내는 수렴 속도가 느리며, 기억 공간을 많이 필요로 하는 단점을 가진다.

본 논문은 강화 학습의 환경에 대한 적응성에 대한 장점을 최대한 살리면서 지연되는 수렴속도와 과중한 기억 공간에 대한 문제를 해결하기 위한 방법론을 연구하였다. 본 논문에서 제안하는 방법은 강화 학습과 같은 행동기반 학습 방법과 기호 학습을 이용하여 규칙을 추출하고 이를 일반화하는 인지 학습 방법을 통합한 방식으로 다음과 2단계 알고리즘이다. 첫 번째 단계는 기호 학습을 통한 규칙 추출로 에이전트의 목적에 맞는 환경 요인을 학습하는 단계이다. 경험들에서 기계학습과 연관 규칙을 이용하여 에이전트 목적에 맞는 환경 요인을 학습한다. 둘째 단계는 수정된 강화 학습으로, 환경에 대한 상호 작용을 통하여 강화 학습을 수행하면서, 기호 학습을 통하여 수행된 학습 결과를 이용한다. 즉, 기호 학습을 통하여 학습된 환경 요인에 대하여 보상값을 예측하여 이를 행위 함수에 적용하는 것이다. 이러한 통합 방식은 행동 기반 학습의 환경 적응성을 인지를 통하여 학습 속도를 향상 시켜 환경 확장성에도 효율적이다.

본 논문에서는 주어진 환경에 대한 적응성뿐 아니라 확장된 환경에서의 효율적인 환경 적응성에 대한 학습 효과를 집중적으로 연구하기 위하여 단일 목적에서의 행동 선택 방법으로 연구 범위를 제한하였다. 이는 단일 목적에서의 행동 선택 효율성을 높여, 이를 기반으로 차후의 연구인 다중 목적을 다루는 행동 선택에 대한 전체적인 학습 효율을 높이기 위한 것이다.

2장에서는 지능형 에이전트에 대한 행동 선택 방법에 대한 기존 연구에 대하여 알아보고, 3장에서는 본 논문에서 제안하는 강화 학습과 인지 학습을 통합한 행동 학습 방법에 대하여 설명한다. 마지막으로 4장에서는 시뮬레이션과 실험에 대한 결과를 통하여 제안 시스템의 우수성을 보이도록 한다.

2. 관련 연구

로봇이나 가상 캐릭터의 연구에서 현재 상황을 인식하고 상황에 맞는 적절한 행동을 선택하는 행동 선택 문제에 대한 많은 연구가 진행되어져 왔다.

첫째, Toby의 연구는 실제 동물의 생태를 연구하는 생태학의 관점에서 연구 된 것으로, 내적 욕구의 만족도를 최대화 시키는 행동을 선택하도록 한다[7]. 이때, Toby는 시스템 디자인 초기에 규칙들을 정의해 놓고 있다. 예를 들어 "Get food"에 대한 행동은 음식을 먹거나, 음식에 가까이 가거나 예전에 먹었던 음식 즉, 기억 속의 음식으로 가까이 가는 것에 의해 만족되며, 음식을 먹는 행위는 씨리얼 먹는 것을 의미한다는 것을 사전에 정의하고 있다. 이러한 Toby의 행동 선택은 시스템 초기에 외부 환경 모델링을 정확하게 수행하며, 이로 인해 동적인 외부 환경에 대한 적응성을 수행하기에 적합하지 않은 단점을 가지게 된다.

둘째, Humphry는 Toby의 정적인 구조를 개선하기 위하여 학습을 이용한다[4]. 여기서 학습은 외부 환경에 적응하기 위해 필요한 지식을 습득하는 과정으로서 Humphry는 강화 학습을 이용한다. Humphry는 전체 환경에 대하여 하나의 강화 학습을 수행하는 것은 불가능 하다고 판단하여 전체 시스템을 여러 가지 목적들로 하위 시스템을 나누고 각각의 시스템에 대하여 강화 학습의 일종은 Q-learning을 이용하여 행동의 패턴을 학습하고 상위 시스템은 W-Learning을 이용하여 학습하였다. 이러한 Humphry방식은 2단계의 학습을 이용함으로 외부 환경에 대한 적응성 면에서 좋은 효과를 보여주나 기본적으로 강화 학습을 이용하여, 학습 속도와 메모리의 부담을 가지게 되며, 이로 인하여 환경에 대한 학습이 비효율적일 수 있다.

본 논문에서 제안한 방식은 Toby의 행동 선택에 대한 효율과 Humphry의 외부 환경에 대한 행동 적응성을 모두 고려하였다.

그래서, 본 논문에서 제안한 통합 방식은 Toby에 대한 정적인 방식을 개선하기 위하여 외부 환경에 대한 학습을 수행하며, Humphry의 방식의 학습 지연과 메모리 사용 부담에 대한 문제를 해결하기 위해서 기호 학습을 이용하여 효과적인 학습을 수행한다. 우리는 이러한 통합된 학습을 통하여 외부 환경의 빠른 적응성을 얻을 수 있었으며, 확장된 외부 환경에서 효율적인 행동 선택을 수행할 수 있었다.

3. 제안 알고리즘

본 논문은 지능형 에이전트의 행동 선택의 학습을 통하여 외부 환경에 대한 적응성 및 확장성을 고려하기 위해서 행동 기반 방식뿐 아니라 인지를 이용한 알고리즘을 제안하였다. 다음 그림은 본 논문에서 제안하는 지능형 에이전트의 구조이다.

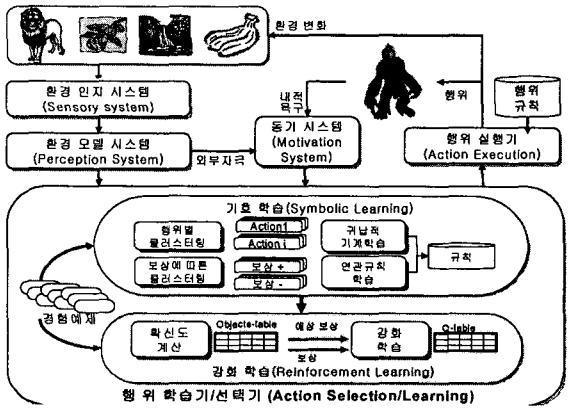

그림 1. 지능형 에이전트 시스템 구조도

지능형 에이전트는 외부 환경을 인식하고 필요한 정보를 추출하는 부분과 이를 기반으로 행동을 선택하고 학습하는 부분과 선택된 행동을 수행하는 부분으로 나눌수 있다. 본 논문은 외부환경을 인식하는 부분과 선택한 행동을 수행하는 부분은 간략화하고, 행동을 선택하고 학습하는 부분을 집중적으로 연구하였다.

본 논문에서 제안하는 지능형 에이전트의 행동 학습 및 선택 알고리즘은 강화 학습과 기호 학습을 통합한 방식으로 다음과 같은 2단계를 수행한다. 첫 번째 단계는 기호 학습을 통한 규칙 추출이다. 입력 예제는 행위별 보상별로 클러스터링을 수행한 후 기계학습과 연관 규칙을 이용하여 규칙을 추출한다. 이 단계에서는 주로 목적에 맞는 오브젝트를 추출하는데 있다. 둘째 단계는 수정된 강화 학습으로, 환경에 대한 상호 작용을 통하여 강화 학습을 수행하면서, 기호 학습을 통하여 수행된 학습 결과를 이용한다. 기존의 강화 학습이 직접 행위에 대한 보상을 받으면서 행위 테이블을 생성하는데, 본 논문에서는 규칙을 이용함으로 예상되는 보상을 이용하여 행위 테이블을 생성하도록 한다.

본 논문에서 제안한 2단계 행동 학습 알고리즘은 기존의 강화 학습만을 고려한 학습 알고리즘에 비하여 학습 속도를 향상 시킬수 있으며, 기호 학습만을 고려한 학습 알고리즘에 비하여 환경에 유연성을 가지고 빠르게 행동을 적용할 수 있는 장점을 가진다. 다음절에서 두단계 행동 학습 알고리즘에 대하여 구체적으로 설명하도록 한다.

3.1 기호 학습(Symbolic Learning)

본 논문에서 기호 학습의 역할은 환경에서 목적에 필요한 오브젝트를 학습하여, 강화 학습에 적용함으로, 전체적인 환경 적응력을 향상시키는 데 있다. 지능형 에이전트가 겪는 경험들을 통하여 규칙을 다양하게 추출하기 위하여 귀납적 기계학습, 연관 규칙을 이용하였다. 이때, 귀납적 기계학습은 단편적인 규칙을 추출하고, 추출된 규칙과 기반 지식을 통하여 일반화를 수행하고, 연관 규칙을 통하여서는 환경의 공간적인 연관성에 관한 규칙을 추출한다.

1) 귀납적 기계학습을 통한 규칙 추출

먼저, 지능형 에이전트의 경험들은 행위별로 클러스터링을 수행하여, 목적에 필요한 행위에 대하여 규칙을 추출한다. 이때 입력 예제들은 보상여부에 따라서 행위에 대한 옳고 그름을 판단 할 수 있다. 이를 통해서 목적에 맞는 환경적인 요인과 목적에 맞지 않는 환경적인 요인을 찾을 수 있다. 본 논문에서는 엔트로피 개념을 활용하는 귀납적 기계학습인 C4.5을 이용한다. C4.5은 Ross Quinlan의 분류모델(Classification Model)로서, 클러스터를 대상으로 각 클러스터를 대표하는 특성(feature)을 발견하고 분석할 수 있다[3]. 본 논문에서는 보상을 받는 행위에 대한 대표적인 특성인 환경요인을 밝히기 위해 정보 이론(Information Theory)에 근거하는 gain값을 사용하는데 gain값을 구하는 식은 다음과 같다[3][6].

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \times Entropy(S_v)$$

$$Entropy(S_v) = -P \oplus Log_2 P \oplus - P \ominus Log_2 P \ominus$$

여기서 S는 전체 집합이며, A는 속성을 나타내는 것으로 여기서는 환경요인을 의미한다. Sv는 속성의 값을 나타내며, P+는 보상을 받은 예제집합을 P-는 보상을 받지 않은 예제 집합을 말한다. 이러한, gain값을 이용한 속성 추출로 현재 목적에 필요한 오브젝트와 목적에 필요하지 않은 오브젝트를 학습할 수 있다.

2) 연관규칙을 통한 규칙 추출

일반적으로, 환경내의 오브젝트간의 관계를 의미하는 공간적 상관관계를 고려하기 위해서 연관 규칙을 이용한다. 본 논문에서는 이러한 공간적 상관 관계를 구현하기 위해 Apriori 알고리즘을 이용하였다. Agrawal et al.이 제안한 Apriori 알고리즘은 비교적 빠른 수행 속도를 가진다[7].

본 논문에서는 복잡도를 줄이기 위하여 보상을 가지는 입력 예제에 한하여 연관 규칙을 적용하도록 하여 목적에 맞는 환경요인에 대한 공간적 상관관계를 파악하여 지능형 에이전트의 행동 선택에 이용하였다. 예를 들어 "Hungry"를 느꼈을 때 "Wheat"가 먹는것임을 인식하고, "Wheat가 보통 Water옆에 있다"는 공간적 상관관계를 파악했다면, 지능형 에이전트가 Water를 인지 했을 때 Wheat가 주변에 있다는 것을 인식하여 행동을 취하도록 하였다. 이러한 공간적 상관 관계를 통하여 지능형 에이전트가 환경에서 보다 유연하게 생활 할 수 있도록 하였다.

3.2 수정된 강화 학습

본 논문에서 제안하는 통합 방식은 지능형 에이전트에 의해서 경험을 통하여 매번 강화 학습을 수행하며, 주기적으로 기호 학습을 이용하여 규칙을 추출 하는 방식이다. 이때, 추출되는 규칙은 강화 학습을 위한 상태 테이블에 영향을 주도록 하였다.

본 논문에서 사용한 강화 학습의 기본 알고리즘은 Q-learning을 사용한다. Q-learning은 문제의 상태 및 행동공간을 Q-table로 만들고 그 상태에 대한 행동의 적합도를 Q-value로 가지며 행동에 따른 결과로 이 값을 갱신함으로서 학습을 하는 방법이다. 이때, 값을 갱신 하는 최적의 행위전략 $\Pi(St)$는 행위 함수로부터 결정 된다[9].

$$\Pi(S_t) = \arg\max_a Q(S_t, A)$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Max_a Q(S_{t+1}, A_t) - Q(S_t, A_t)]$$

여기서 St는 현재 상태를, At는 현재 상태에서 취한 행위를, St+1는 현재 상태에서 행위를 취했을 때 일어나는 다음 상태를, R은 직접적인 보상을 말한다. 이때, Q-learning의 행위 전략은 현재 생태(St)에서 가장 값이 높은 행위(A)를 선택하는 것이며 행위 함수는 위 식처럼 갱신된다.

이처럼 강화 학습은 계속 되는 경험을 통하여 행위 값을 수정하는 Q-table을 유지하고, 이를 기반으로 행위를 선택하게 된다. 하지만 Q-table을 이용하여 지능형 에이전트의 적절한 행위를 학습하는데는 많은 시간이 걸리게 된다. 그래서 본 논문은 기호 학습을 통하여 습득한 규칙을 이용하여 행위 함수를 수정 하도록 설계하였다.

기존의 강화 학습은 Q-table만을 이용하는데, 본 논문은 오브젝트-table을 이용하여 환경 요인인 오브젝트들에 대한 학습을 수행하도록 하였다. 오브젝트-table은 각 오브젝트가 각 목적에 얼마나 적합한지에 대한 확신도를 가지고 있다. 예를들어, Hunger라는 문제를 해결해야하는 시스템에서 Wheat가 얼마나 적합한지에 대한 값을 나타내는 것이다. 이때, 오브젝트-table의 확신도는 기호 학습을 통하여 계산하게 된다. 이 확신도를 이용한 행위 함수는 상황에 따라 다른 보상 함수(Func(Rt+1))를 가지게 된다.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[Func(R_{t+1}) + \gamma MaxQ(S_t, A_t) - Q(S_t, A_t)]$$

$$Func(R_t) \leftarrow \text{If Rt+1=0 return Pred(Rt+1) else return Rt+1;}$$

Q-table내에 보상값이 있다면 그것을 이용하고 보상값이 없으면서 목적에 대한 확신도를 가지는 오브젝트가 있다면 예상 보상값을 가지게 된다.

다음은 본 논문에서 제안한 강화 학습 알고리즘 이다.

1. Q-테이블 및 각 파라미터(a, r) 등을 알맞게 초기화 한다.
2. 다음 내용을 반복한다.
   1) Q-테이블로부터 행위를 결정한다.
      이때 임의의 비율로 랜덤 행위를 만들어 주어야 한다.
   2) 행위에 대한 다음 상태와 보상을 얻는다. 이때 오브젝트 테이블의 확신도를 이용하여 예상 보상값을 계산한다.
   3) 현재 상태의 행위값을 갱신 한다.
   4) 행위 책략을 갱신한다.
   5) 주기적으로 규칙 추출 및 일반화를 수행하고 결과를 오브젝트 테이블에 반영한다.
      - 오브젝트 테이블의 확신도 값과 보상값을 갱신한다.

이처럼 본 논문에서 제안하는 지능형 에이전트의 2단계 행동 학습 및 선택 알고리즘은 기호 학습을 통하여 오브젝트를 학습하고, 학습한 내용을 Q-table에 적용함으로 학습의 속도를 향상시켜, 외부 환경의 빠른 적응성과 확장성을 가지도록 하였다.

4. 시뮬레이션 및 실험

본 논문에서 제안한 행동 선택 학습 알고리즘의 타당성과 성능을 검증하기 위해서 AI평원 시뮬레이션에서 실험 하도록 한며, 제안한 알고리즘에 대한 평가는 학습속도 측면을 보도록 한다.

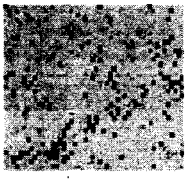

그림 2. AI평원

다음은 지능형 에이전트를 실험 하기 위한 자바로 구현한 40*40의 AI평원 환경을 보여주고 있다. 이 환경은 Jackson Pauls가 생성한 아프리카 평원을 확장한 것이다[2]. 이 환경에는 물과 늪이 있고 주식으로 <wheat, rice, potato>가 있고, 나무로 <oak, pine>이 있고, 꽃으로 <rose, azalea, lily>가 있다. 이 환경은 기본적으로 격자(grid world)세계로 이루어 졌다. 이 세계를 살아가는 지능형 에이전트는 내부적인 욕구로 < hunger>를 가지고 있으며, 시계 범위는 2로 제한하였고, 매츠텝 마다 hunger의 level이 1씩 올라가며, 15가 되면 죽게 설정 하였다. 지능형 에이전트가 할 수 있는 행동은 먹는 것과 이동하는것(8방향)으로 제한 하였고, 지능형 에이전트가 먹을수 있는 것을 먹었을 때 먹은 양에 따라 보상을 받도록 하였다.

다음은 AI평원에서 지능형 에이전트의 시간에 따른 평균 수명률을 보여주고 있다. 점선은 일반 Q-learning을 수행한 것이고 실선은 추론된 규칙과 Q-learning을 이용한 것이다.

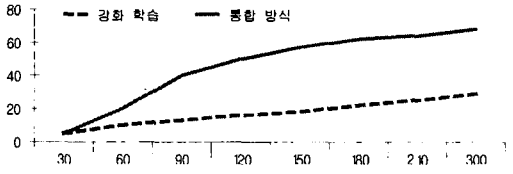

그림 3. 학습 속도 실험 평가

이 실험을 통하여 기존의 Q-learning에 비하여 추론된 규칙와 연관 규칙, 일반화를 이용한 Q-learning이 빠르게 생명을 안정화 시켜감을 알 수 있었다. 즉, 본 논문에서 제안한 방식이 에이전트에 대한 학습을 보다 빠르게 수행 시켜 외부 환경에 빠르게 적응 하고 있음을 확인 할 수 있었다.

5. 결론 및 향후 연구

인간은 자신의 욕구를 느끼고 주어진 환경을 인식하여 살아가기 위한 최선의 행동을 끊임없이 선택한다. 본 연구는 이러한 인간의 행동 학습과 선택 대한 메카니즘을 인공지능 측면에서 연구하여 효율적인 지능형 에이전트를 만들고자 하였다. 본 논문에서 제안하는 방법은 강화 학습과 같은 행동기반 학습 방법과 귀납적 기계학습, 연관규칙을 이용한 인지 학습 방법을 통합한 방식이다. 이방법을 통해 외부 환경에 대한 빠른 적응성을 보였으며, 확장된 환경에서 이전의 학습 결과를 효과적으로 이용 할 수 있어, 변화하는 환경에 빠르게 적응 할 수 있었다. 본 논문에서 제안한 방식은 AI평원에서의 시뮬레이션 실험을 통해 우수성을 확인 할 수 있었다.

본 논문은 지능형 에이전트에서 단일 목적에서의 외부 환경의 적응성과 확장성을 위한 학습에 대한 연구를 수행하였다. 차후 연구에서는 이러한 학습 방법을 이용한 단일 학습의 그룹인 다중 목적들간의 학습에 대한 연구를 수행하고자 한다. 현재, 여러 가지 목적을 동시에 만족 시키는 행동을 학습하는 방법에 대한 연구가 진행되고 있다. 이와 더불어, 감정을 이용한 행동 선택과 행동 선택에 따른 감정의 변화에 대한 연구를 수행하고자 한다. 이와 같은 연구는 지능형 에이전트가 인간과 같은 학습을 수행하며, 감정을 표현함으로 차세대의 인간 친화적인 인터페이스로서의 역할을 수행 할 수 있을 것이라고 예상된다.

6. 참고문헌

[1] Carlos Gershenson , "Philosophical Ideas on the Simulation of Social Behaviour", Journal of Artificial Societies and Social Simulation vol. 5, no. 3, 2002.
[2] Jackson pauls, Pigs and People, 4th year report, 2001.
[3] J R. Quinlan, "C4.5 Programs for Machine Learning", San Mateo, CA:Morgan, Kaufaman, 1993.
[4] Mark Humphrys, "Action selection methods using reinforcement learning", University of Cambridge, 1997.
[5] Pattie Maes, "Modeling adaptive autonomous agents", Artificial Life Journal, vol.1, no.1~2, pp.135~162, 1994.
[6] T. Mitchell, "Machine Learning", McGraw Hill, 1997.
[7] T. Tyrrell, "Computational Mechanism Action Selection", Ph.D. Thesis, University of Edinburg, 1993.
[8] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In *Proceedings of the 20th VLDB Conference,* Santiago, Chile, Sept., 1994.
[9] R. Sutton, A. Barto, Reinforcement Learning, MIT Press, 1997.
[10] S. Wermter and R. Sun, (eds.) Hybrid Neural Systems. Springer-Verlag, Heidelberg. 2000.