Chapter 01

유전 알고리즘의 개괄

*** 학습목표**

- 유전 알고리즘의 역사를 개괄적으로 살펴본다.
- 유전 알고리즘의 구조, 연산들을 가벼운 수준에서 이해한다.
- 유전 알고리즘이 모든 문제에 적합한 기법은 아님을 이해한다.

Preview

태초에 DNA가 있었다.

유전 알고리즘은 컴퓨터 과학에서 가장 최근에 세력화가 이루어진 연구 분야 중 하나다. 1960년대에 태동한 이래 1990년대 이후 폭발적인 연구 결과들을 양산해내고 있다. 이같은 현실과 어울리지 않게도 유전 알고리즘의 대부인 존 홀랜드는 사실상 미국 최초의 컴퓨터 과학 분야 박사학위 취득자다.

이 장에서는 20세기 후반부터 본격적으로 조명을 받고 있는 진화적 프로세스를 기초로 한 문제 해결 방법론인 유전 알고리즘의 역사를 간략히 소개하고, 2장 이후에 배울 내용들을 간략히 스케치한다. 우선 유전 알고리즘의 전형적인 구조, 문제의 해를 유전 알고리즘에서 표현하는 방법, 두 해를 결합하여 새로운 해를 만드는 방법, 세대가 지남에 따라 해집단의 구성원을 바꾸는 방법, 유전 알고리즘의 작동 과정에서 중요한 역할을 하는 부분 패턴인 스키마 등을 가벼운 수준으로 소개한다. 그런 다음, 유전 알고리즘이 모든 문제에 적합한 기법이 아닌 이유도 설명한다.

Section

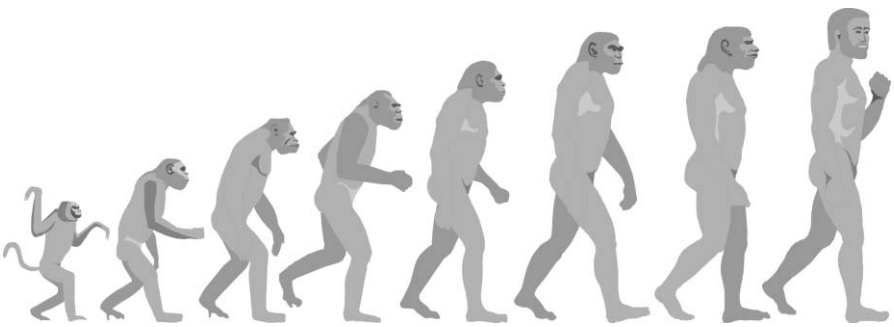
01 진화

인간의 탄생은 정보로부터 시작된다. 부모의 DNA가 혼합되어 길이가 약 30억인 염기 서열이 만들어지는데 이것이 한 인간의 시작을 규정짓는 거의 전부다. 염기는 A, C, T, G 중의 하나이므로 이를 0, 1, 2, 3으로 대응시키면 이것은 약 30억 자리의 4진수에 해당된다. 이 정보로부터 인간의 신체가 만들어진다. 동일한 DNA라 하더라도 성장 환경에 따라 차이를 보일 수 있지만 지능이나 심미적 성향 등은 타고난 DNA를 극복할 수 없는 분명한 한계가 있다. 한 인간은 약 30억 자리의 4진수로부터 시작되는 것이다. 생태계에서 진화는 쉴 새 없이 새로운 4진수들을 만들어 낸다. 생태계는 4진수들의 경쟁의 장이다.

필자가 미국에서 공부하던 1990년대 초반에 어느 잡지에서 한 컬럼니스트가 "다음 세기 초반 몇십 년간의 주제는 진화"라고 쓴 것을 본 기억이 있다. 일반적으로 생물학의 한 분야로만 인식되던 진화가 20세기 후반부터 새로이 많은 지성들의 주목을 받고 있다. 1859년 찰스 다윈에 의해 공식적으로 발표된 이후 진화는 창조론으로 대표되는 종교적 신념과 꾸준히 대립해왔다. 칼 마르크스는 진화론의 생존 경쟁이 인류사의 계급 투쟁을 자연계에 대입해 놓은 것이라고 하며 자신의 계급 투쟁론을 정당화하는 데 사용했는데, 이 사실이 진화론의 입지를 더욱 좁혀 놓았다. 1925년 미국 테네시주에는 학생들에게 진화론을 가르친 교사에게 100불의 벌금이 부과된 망키 추라이얼이라는 유명한 재판도 있었다. 1980년대 말에는 미국의 루이지애나 대법원에서 창조 과학을 진화론과 동일한 수준으로 가르쳐야 한다는 주법이 청문회를 거쳐 철폐된 사건이 있었다. 1999년에 미국의 캔사스주 교육위원회에서는 초중등 교육을 위한 과학 교육 표준에서 진화론과 빅뱅 등 지구의 연대기적 나이를 암시하는 모든 내용을 삭제하기로 결정된 사실도 있다. 2004년 미국 펜실바니아주의 작은 도시인 도버의 교육위원회에서 중고등학교 과학 시간에 학생들에게 "진화론은 완벽하지 않으며 생명의 기원을 설명하는 다른 가설인 지적 설계론(창조론)도 있다"는 요지의 1분짜리 성명서를 의무적으로 읽어주도록 결의했다가 법원에서 위헌 판결을 받아 좌절된 바 있다.

반세기 전까지만 해도 로마 교황청은 진화에 대한 공식적인 논의를 허용조차 않을 정도로 진화에 대해 분명한 반대의 입장을 견지했으나, 1950년의 한 교지를 통해 "진화에 대한 논의는 허용하나 다만 공산주의자들이 신을 깎아 내리기 위해 진화를 악용하는 것은 경계해야 한다"고 진화에 대한 관점을 다소 누그러뜨렸다. 이로부터 반세기가 지난 1996년 10월 4일, 로이터 통신에는 교황청의 전격적 화해 선언에 해당하는 기사가 타전되었다. 전문에 의하면 당시의 교황 요한 바오로 2세는 "지난 1950년의 교지 이후 새로이 발견된 지식들은 이제 진화론이 더 이상 가설에 머무를 수 없음을 인정하게 만든다"고 말하였다. 다만 "인간의 육체가 그 이전에 존재하던 생명체에 그 근원을 갖고 있다 하여도 그 영혼(spiritual soul)은 신에 의해 직접적으로 만들어진 것이다"라고 부연함으로써 진화론과의 화해를 택했다.

필자는 "창조냐 진화냐"하는 해묵은 논쟁에 참여할 의도는 없다. 다만, 진화는 이제 옛날의 비종교적 이데올로기의 이미지에서 탈피하였고 현대의 과학 및 사회, 경제 시스템의 다양한 현상을 설명하거나, 어려운 문제를 해결하는 원리로서 새로운 조명을 받고 있다는 사실을 말하고 싶은 것뿐이다. 유전 알고리즘은 영어로 Genetic Algorithm, 줄여서 GA라고 하는데 진화의 원리를 문제 풀이 또는 모의실험에 이용하는 연구 방법인 "진화 연산"(Evolutionary Computation)의 대표적인 한 분야다.



Section

2 유전 알고리즘의 약사

유전 알고리즘은 진화의 원리를 문제 해결에 이용하는 대표적인 방법론 중 하나다. 유전 알고리즘의 대부는 유명한 존 홀랜드(John Holland)다. 그러나 진화의 원리를 이용하는 문제 해결 방법은 홀랜드가 최초는 아니다. 1960년대에 독일의 Rechenberg(1965)가 진화 전략(Evolution Strategy)을 제안하였고, Fogel, Owens, Walsh(1966)에 의해 진화 프로그래밍이 제안되었다. 이에 앞서 몇몇 연구자들이 진화에 근거한 접근법을 사용한 적이 있다 [Box, 1957; Friedman, 1959; Bremermann, 1962]. 적지 않은 연구 결과들이 홀랜드와 동시대 또는 그 이전에 있었으나 홀랜드가 진화 연산의 대부격으로 인정을 받는 데는 이유가 있다. 초기의 Rechenberg는 단 한 개의 해를 변형시켜가는 방법을 사용하고 유전 알고리즘의 주 연산인 교차 연산은 사용하지 않았다. Fogel 등은 교차 연산 없이 변이만을 사용하고 아직도 그 전통을 고수하고 있다(해집단, 교차, 변이 등의 용어에 대해서는 이 장에서 다룬다). 집단에 근거하고 교차와 변이를 포함한 골격을 완성한 것은 홀랜드의 공로고, 더욱 중요한 것은 1975년 그의 역사적인 저서 『Adaptation in Natural and Artificial Systems』를 세상에 발표하여 유전 알고리즘의 이론적 기반을 다졌다는 점이다.

자신의 연구에 대한 홍보에 미숙한 홀랜드의 성향이 유전 알고리즘의 발전 속도를 더디게 했다고 말하는 사람도 많다. 그러나 홀랜드는 노벨 물리학상을 받은 머레이 겔만(Murray Gell-Mann)과 필 앤더슨(Phil Anderson), 노벨 경제학상을 받은 케네스 애로우(Kenneth Arrow) 등이 주축이 되어 복잡성 과학의 학제간 연구를 위하여 1984년 결집한 산타페 연구소에 참가하면서, 이 연구소의 연구 방향을 "복잡계"에서 "적응적 복잡계"로 선회하도록 할 정도로 영향을 미쳤고, 유전 알고리즘의 황금기를 본격적으로 열었다.

1985년에는 최초의 유전 알고리즘 전문 학회인 제 1회 International Conference on Genetic Algorithms(ICGA)가 개최되었다. 1990년대에 많은 주목을 받은 인공 생명(Artificial Life)[Langton, 1989]의 주된 도구가 유전 알고리즘이란 사실도 이 분야의 발전을 가속화시켰다. 이제는 컴퓨터과학, 물리학, 생물학, 화학, 경제학, 각종 공학 등의 연

구자들이 공동으로 관심을 갖는 연구 주제로 확립되었다. 1997년에는 전기 · 전자 · 컴퓨터 분야의 가장 대표적인 학술지인 IEEE Transactions에 유전 알고리즘으로 대표되는 진화 연산 전문 트랜잭션이 개설되어 이 분야의 역사에 한 획을 그었다. 현재 IEEE Transactions on Evolutionary Computation은 IEEE Transactions 중에서는 물론이고 컴퓨터 과학 관련 학술지들 중 가장 인용지수(impact factor)가 높은 것 중의 하나로 자리 잡았다. 격년으로 개최되던 ICGA는 1999년 Genetic Programming Conference와 합병하여 현재는 Genetic and Evolutionary Computation Conference(GECCO)란 이름으로 매년 개최되고 있다.

스키마: 1.6절 참고

홀랜드의 제자로 1980년대 중반 이후 유전 알고리즘 분야에서 가장 왕성한 저술 활동을 보인 사람은 데이빗 골드버그(David Goldberg)다. 그는 Fogel, Owens, Walsh 등의 초기 "진화 프로그래밍" 연구자들의 연구가 학계에서 인정을 받지 못하고 거부당한 나머지 스키마(1.6절에서 설명 예정) 이론에 기반한 유전 알고리즘까지 유사하게 모호한 경향으로 취급받게 만들어 6, 70년대에 유전 알고리즘에 대한 회의론을 불러일으킨 한 중요한 원인으로 보고 있다. 이 의견을 그의 저서 『Genetic Algorithms in Search, Optimization, and Machine Learning』(Addison-Wesley, 1989)에서 언급했을 정도다.

Section

03 유전 알고리즘의 기본 용어들

생물학에서 유전 물질은 DNA, 즉 염색체(chromosome)다. 우리 인간의 세포 하나마다 DNA 한 쌍이 있어 각 DNA는 30억 개 가량의 염기로 되어 있고, 이들은 4만 개 내외의 유전자를 포함하고 있는 것으로 추정되고 있다. 개체들은 교차에 의해 염색체를 부분 결합하고 돌연변이에 의해 미소하게 변화된 새로운 염색체를 가진 새로운 개체를 만들어 낸다. 개체는 환경에 적응하기 유리한 정도에 따라 선택적으로 (경쟁적으로) 번성한다. 유전 알고리즘의 기본 구조는 이러한 생물의 진화 과정을 문제 해결 과정으로 옮겨 놓은 것이다. 문제 해결상의 임의의 해를 유전 알고리즘이 이해하는 형태로 표현한 것을 염색체라 부른다. 자연계의 가변적이고 제한이 없는 개체들의 집단 대신, 유전 알고리즘에서는 대부분의 경우 정해진 수의 염색체 집단을 운영하는데 이를 해집단(population)이라 한다. 즉, 유전 알고리즘은 복수 개의 해를 유지하면서 운용된다.

염색체상의 각 인자를 유전자(gene)라 한다. 생물학에서는 많은 수의 염기가 모여 유전자를 형성하는데 유전 알고리즘에서는 유전자가 최소 단위다. 즉, 염기와 같은 유전자의 하위 개념이 없다. 생물학에서 유전자형(genotype)은 보이는 그대로의 유전자 조합을 뜻하고, 표현형(phenotype)은 유전자형과 관계되어 관찰되는 형질을 의미한다. 유전 알고리즘에서도 비슷하게 염색체 그 자체를 유전자형이라 하고, 이와 대응되는 해의 성격이나 품질 등을 표현형이라 한다.

Section

4 유전 알고리즘의 전형적인 구조

[알고리즘 1-1]은 전형적인 유전 알고리즘의 구조다. 유전 알고리즘은 대부분 정해진 수의 해로 구성되는 해집단을 갖는다. 이 알고리즘에서 해집단의 해의 수는 n이다. 먼저 n개의 해를 임의로 생성한다. 이 해집단으로부터 k개의 새로운 해를 만들어 내는데 각각의 해는 선택(selection), 교차(crossover), 변이(mutation)의 단계를 거쳐 만들어진다. 이렇게 만들어진 k개의 해는 해집단 내의 k개의 해와 대치된다. 이러한 과정을 임의의 정지 조건이 만족될 때까지 수행한 후 해집단에 남은 해 중 가장 좋은 해를 답으로 삼는다.

상수 k는 해집단이 한번에 얼마나 많이 대치되느냐를 결정하는데, k/n를 세대차(generation gap)라 한다. 세대차가 1에 가까운, 즉 절대 다수의 해가 대치되는 경우를 "세대형 유전 알고리즘(generational GA)"이라 하고, 세대차가 $1/n$에 가까운, 즉 새로운 해가 생기자마자 해집단에 넣어주는 방식을 "안정 상태 유전 알고리즘(steady-state GA)"이라 한다. 이 두 유형의 유전 알고리즘들간에는 서로 장단점이 있으나, 대체로 안정 상태 유전 알고리즘이 해집단을 빨리 수렴시키는 경향이 있는 대신 설익은 수렴(premature convergence)의 가능성도 더 크다. 어느 한 유형을 사용할 때는 해당 유형의 결점들을 보완하는 다양한 기술이 있다. 위 작업들의 세부 사항에 대해서는 3장에서 상세히 설명한다. 유전 알고리즘은 다양한 형태와 변형이 있으므로 모든 유전 알고리즘이 이 구조와 일치하는 것은 아니다.

연산들 중 선택은 교차를 위해 해집단에서 임의의 해를 선택하는 연산이다. 여기서는 우수한 해에게 선택될 확률을 높게 준다는 것 정도만 이야기하고 상세한 설명과 변형은 3.1절에서 다룬다. 이렇게 선택된 해를 부모해(parent)라 한다. 교차는 두 개의 부모해로부터 자식해(offspring)를 하나 만들어 내는 연산이다. 교차는 유전 알고리즘의 대표적 연산으로서 유전 알고리즘의 성능에 지대한 영향을 미친다.

[알고리즘 1-1] 유전 알고리즘의 전형적 구조

n개의 초기 염색체 생성;
repeat {
    for $i = 1$ to k {
        두 염색체 p_1, p_2 선택;
        $offspring_i = crossover(p_1, p_2)$;
        $offspring_i = mutation(offspring_i)$;
    }
    $offspring_1$, ..., $offspring_k$를 population내의 k개의 염색체와 대치;
} **until**(정지 조건 만족);
남은 염색체 중 최상의 염색체를 **return**;

과거에는 교차 연산이 어떠한 것이라는 공통의 특징들이 있었는데 요즘은 그런 식으로 설명하기는 어렵다. 복수 개의 해를 결합하여 하나를 만든다는 점이 거의 유일한 공통점이라 할 정도로 다양해졌기 때문이다. 변이는 해를 임의로 변형시키는 연산이다. 교차가 두 부모해에 있는 속성들을 부분적으로 이용하는 역할을 하는 반면, 변이는 부모해에 없는 속성을 도입하여 해의 다양성을 높이는 역할을 한다.

교차의 원리는 1.7절 참고, 다양한 교차 연산은 3.2절 참고

변이의 원리는 1.8절과 3.3절 참고

유전 알고리즘이 정지하기 위한 조건은 다양하게 줄 수 있다. 가장 대표적인 두 가지는 [알고리즘 1-1]의 **repeat-until** 루프를 일정 횟수만큼 수행한 다음 정지시키는 방법과 해집단에 있는 해들의 다양성이 어느 정도 이하로 떨어지는 시점에 정지시키는 방법이다. 다양성이 떨어지는 것을 판단하기 위해서는 해집단 내의 염색체들 중 대부분이(예를 들면 70%) 똑같은지를 확인하는 경우가 일반적이다. **repeat-until** 루프를 일정 횟수만큼 수행한 다음 정지하도록 설계된 유전 알고리즘의 경우라도 그 정도면 해들이 어느 정도 수렴할 것이라는 경험적 짐작이 있어야 한다. 많은 유전 알고리즘 설계자들이 문제의 복잡도와 유전 알고리즘의 작동 경향에 대한 경험적 직관없이 이것을 임의로 설정하는데 이것은 매우 위험하다.