유전 알고리즘과 군집 분석을 이용한 확률적 시뮬레이션 최적화 기법

이동훈(국방과학연구소 제 2 연구개발본부 수중탐지체계부)
허성필(해군사관학교 경영과학과)

Abstract

유전 알고리즘은 전통적인 등반 알고리즘을 이용하여 구하기 어려웠던 최적화 문제를 해결하기 위한 강인한 (Robust) 탐색 기법이다. 특히 목적함수가 (1)여러 개의 국부 최대치를 가지거나 (2)수학적으로 표현이 불가능하거나 어렵거나 (3)목적함수에 교란항이 섞여 있을 경우도 우수한 탐색 능력을 갖는 것으로 알려져 있다.

본 논문에서는 군집성 분석(cluster analysis)을 이용하여 군집화함으로써 유전 알고리즘을 이용하여 나타나는 다양한 해집합을 형성하는 개체군을 그룹화하고, 각 군집에 부여된 군집 적합도에 따라서 최적해를 구함으로써 최적값에 근접시킬 수 있는 탐색 알고리즘을 제안하였으며, 시뮬레이션의 출력이 특정한 테스트 함수의 형태로 나타난다고 가정한 경우에 확률적으로 나타나는 시뮬레이션 모델의 출력을 최대화하는 문제에 대하여 적용하고 분석하였다.

1. 서론

컴퓨터를 이용한 시뮬레이션은 시간적으로 또는 기술적으로 구현하기 어려운 해석적인 방법에 대한 대안이다. 모델 구축의 초기 단계에는 시뮬레이션의 반응면이 단일 피크(peak)여부와 미분가능성 여부를 판단하기 어려우며, 또한 시뮬레이션의 출력이 성공 실패로 나타날 경우의 반응면은 베르누이 시행을 따르는 확률변수가 되어서 최적화 문제로서 전통적인 등반 알고리즘의 적용이 어려워 진다. 이러한 문제에 있어서 유전 알고리즘은 인공지능적 기법 중의 하나로서 다양한 후보 해를 제공하고 후보 해들간의 정보 교환으로 최적해에 도달하는 확률을 높이는 강인한 탐색 알고리즘이다.[1,2]

(1)반응면의 형태를 예상할 수 없고 (2)출력의 형태가 베르누이 확률변수로 나타나는 (3)시뮬레이션 프로그램을 이용하여, 최적해를 찾는 문제를 해결하는 문제는 최종적 해의 정확성 뿐만이 아니라 해에 근접해가는 속도가 중요한 요소이다. 허성필[3]은 이러한 문제를 해결하기 위하여 단순 유전 알고리즘(simple genetic algorithm)을 변형한 방법을 제안하였으며 매개변수에 따른 탐색 능력의 변화를 몬테칼로 시뮬레이션을 통하여 분석함으로써 확률적 시뮬레이션 문제에서 탐색 전략을 어떻게 세워야 할 것인지를 분석하였다.

본 논문에서는 변형된 단순 유전 알고리즘의 다양한 해집합의 군집성을 분석함으로써 적은 시행횟수로 실제해에 좀 더 근접시키는 알고리즘을 제안하고자 한다.

2. 성공확률 최대화를 위한 유전 알고리즘

제안된 알고리즘은 성공 여부로 출력이 나타나는 시뮬레이션의 성공확률을 최대화시키는 입력조건을 찾는 문제를 해결하기 위한 것이다.

가. 적용 유전 알고리즘

■ 단계 1.
초기화 : 세대 k = 0, 랜덤으로 popsize 개의 개체를 선택한다.

$$X_i = (x_{0i}, x_{1i}, x_{2i}, x_{3i}, x_{4i}), i = 1,2\ldots, popsize$$

■ 단계 2.
성공확률 추정 시뮬레이션 수행 및 적합도 계산 : 각 개체에 대하여 n 회의 시뮬레이션을 수행하여 누적 성공횟수 M_{ik}와 누적 시행횟수 N_{ik}를 구하며 이를 이용하여 식 (1)의 성공확률의 추정치를 구한다. 이 때 k 번째 개체가 이전 세대에서의 개체가 그대로 유전된 경우 M_{ik}와 N_{ik}는 이전 세대의 누적 시행횟수 M_{ik-1}와 성공횟수 N_{ik-1}에 현재의 시행횟수와 성공횟수를 추가할 수 있도록 알고리즘을 구성한다. 이렇게 함으로써 반복 횟수가 누적된 우수한 개체는 실제 성공확률에서 벗어날 확률이 작아지므로 우연히 도태될 가능성이 작아지며, 우연히 선택된 열등한 개체는 다음 세대에서 사라질 가능성이 커진다.

성공확률 추정치 : $p_{ik} = M_{ik} / N_{ik}$ ----------(1)
누적성공횟수 : $M_{ik} = M_{ik-1} + m_{ik}$
누적시행횟수 : $N_{ik} = N_{ik-1} + n$

■ 단계 3.
다음 세대의 선택 : 세대 수 k 를 1 증가시킨다. 적합도 상위 (1-pcross)*100%의 개체는 다음 세대로 변화없이 그대로 보낸다.

■ 단계 4.
pcross*100%개의 개체는 식 (2)의 확률에 따라서 선택하고 교차 변이를 수행하고 pmutation 의 확률에 따라서 돌연변이를 수행함으로써 새로운 개체를 발생시킨다.

----------(2)

이 때의 누적성공횟수 N_{ik} 와 누적시행횟수 M_{ik} 를 모두 0 으로 놓는다.

■ 단계 5.
단계 2 - 단계 4 를 maxgen 만큼 반복 수행 후 최종적으로 선택된 개체군을 이용하여 군집성 분석을 실시하여 최적해를 구한다.

나. 적합도 함수

1 회의 결과가 성공 또는 실패의 형태로 나타나는 시뮬레이션은 성공확률 p 인 베르누이 시행으로 볼 수 있으며, N 회의 베르누이 시행에서의 성공횟수 M 의 분포는 식 (3)과 같으며 근사적으로는 식 (4)와 같다.

$$M = Np \sim Binom(N, p) \quad \text{----------(3)}$$

$$\text{----------(4)}$$

N 이 유한하면 성공확률의 추정치 p_{ik} 는 근사적으로 실재성공확률 p_{ik} 를 의 오차로 추정하게 되며, 따라서 유전적 선택시 낮은 성공확률을 갖는 개체가 채택되거나 높은 성공확률을 갖는 개체가 탈락될 수 있게 한다. 따라서 N 을 조정하여 적절한 분산을 가지도록 함으로써 우수한 개체가 도태되는 확률을 조절해야 우수한 탐색 능력을 가질 수 있다.
성공확률 p 의 α*100% 신뢰 구간의 하한치는 식 (5)와 같이 구해지며, 이 값를 각 개체의 적합도로 정의한다.

$$F = p - 1/\sqrt{4N} \quad \text{----------(5)}$$

$p = M/N$ 대신에 F 를 이용하는 것은 특정 개체의 $p = M/N$ 이 조금 작더라도 N 이 큰 개체는 여러 세대를 거쳐서 유전된 개체이므로 높은 적합도를 주는 것이 바람직하기 때문이다. T 를 너무 크게 설정할 경우 처음 들어 오는 개체는 N 이 작게 되므로 우수한 개체라도 기존의 열등한 개체를 도태시키고 최종적인 해집합의 하나로 남기가 어려우며, 반대로 N 이 너무 작을 경우 우수한 개체가 열등한 개체가 우연히 얻은 높은 적합도 때문에 도태될 확률이 높아 질 수 있다. 따라서 적절한 T 를 설정함으로써 효과적인 탐색 전략을 수립하여야 한다.

다. 군집성 분석 및 군집 적합도

유전 알고리즘의 특징 중의 하나는 한 개의 최적해를 제시하는 것이 아니고 다양한 복수의 해를 동시에 제공해 주는 것을 장점으로 가지고 있다. 그러나 이러한 다양한 해가 모두 국부 피크치라고 보기에는 무리가 있고 상당수가 국부 피크치의 근방에 위치한 개체라고 볼 수 있다. 이럴 경우 이들 국부 피크치 주의 의 값 중 하나를 택하여 최적해로 선택하는 것보다 이들의 평균을 이용하는 것이 보다 효과적일 수 있다. 군집 분석(cluster analysis)[4]은 그룹의 수나 구조에 대한 아무런 가정이 없이 유사성(similarity measure)만으로 각 관측치(개체)들을 임의의 그룹에 할당하는 기법으로 하나의 국부 피크치를 중심으로 갖는 개체들을 그룹화하는 알고리즘으로 적용이 가능하다. 군집성 분석 및 군집 적합도 부여 절차는 다음과 같다.

■ 단계 1.
선정된 개체들을 이용하여 n_cluster 개의 군집으로 나눈다. 개체군을 n_cluster 개의 군집으로 나누는 절차는 다음과 같다.
1) 군집수 n_cluster 를 2 로 놓는다.
2) 각 개체들은 초기 군집으로 나눈다.
3) 각 군집의 대표치(centroid) C_i 를 구한다.

$$C_i = (c_{i0}, c_{i1}, \ldots, c_{ip}), i = 1, 2, \ldots, n\_cluster$$

여기서, c_{ijk} 는 i 번째 군집의 k 번째 개체의 j 번째 변수의 값, N_{ik} 는 i 번째 군집의 k 번째 군집의 개체의 수, η_i 는 i 번째 군집의 개체의 수이다.
4) 대표치에 가장 인접한 군집으로 각 개체를 재할당한다. 인접한 정도는 대표치와 각 개체간의 거리로 계산한다.
5) 3)와 4)를 반복하여 대표치와 개체간의 거리의 합이 최소가 될 때까지 반복한다.
6) 군집 대표치로부터의 거리가 R 이상인 개체가 존재할 경우 n_cluster 를 하나 증가시키고 2)에서 5)를 반복 수행한다. 최대 거리가 R 이하일 경우 군집 분석을 중단한다.

■ 단계 2.
단계 군집의 평균을 이용하여 군집화된 개체를 구하며 식 (6)과 같이 군집 적합도를 구한다.

$$F_{cluster}(i) = p_{cluster}(i) - 3/\sqrt{4N_{cluster}(i)} \quad \text{---(6)}$$

여기서,

$$p_{cluster}(i) = M_{cluster}(i)/N_{cluster}(i)$$

■ 단계 3.
최대의 군집 적합도를 갖는 군집화된 개체를 최적값으로 선정한다.

3. 테스트 함수에 대한 알고리즘 적용

제안된 알고리즘을 이용하여 식 (7)과 같이 주어진 테스트 함수에 대하여 얼마나 효과적으로 최적해를 탐색할 수 있는지를 모의실험을 통하여 평가하였다.

가. 테스트 함수

반응면의 표면이 여러 개의 국부 최대치를 갖거나 확률 변수의 형태를 갖는 경우의 탐색 알고리즘의 효과도 비교를 위한 테스트 함수는 여러 가지가 알려져 있으나[2] 여기서는 식 (7)과 같은 테스트 함수를 정

의하고 유전 알고리즘의 효과도를 알아 보았다. 이 함수의 형태는 [그림 1]과 같이 2^{g+1}개의 국부 최대치를 가지는 형태의 테스트 함수이다.

$$\sum_{i=0} \sin(x_i) \ldots, \quad 0 \le x_i < 180 \qquad (7)$$

나. 알고리즘 적용성 실험 계획

유전 알고리즘의 최적값을 찾는 효율성을 평가하기 위하여 [표 1]과 같은 매개변수의 조합에 대하여 시뮬레이션을 실시하였으며 각 세대별로 군집 분석을 실시하여 군집을 나누고 각 군집에 군집 적합도를 부여하고 최적의 군집을 최적치로 선정하였다.

[그림 1] 테스트 함수의 형태(g=1)

[표 1] 적용 유전 알고리즘의 매개변수 값

매개 변수	값
개체군의 크기	200
최대세대수	100
개체당 시행의 수	6
교차 확률	0.6
돌연변이 확률	0.003
T	0.5
R	10

알고리즘의 효율성을 평가하는 측도는 식 (8)과 같은 개념의 알고 있는 최적해와 추정된 최적해 간의 거리를 이용한다.

---(8)

여기서 $\bar{\mu}_i$ 와 μ_i 는 각각 알고 있는 최적해과 알고리즘에 의해서 추정된 최적해이다.

다. 결과 분석

[그림 2]는 단순히 유전 알고리즘을 적용하여 최대값을 찾아 나가는 과정을 보여주는 그림으로 세대의 진행에 따라서 일정한 값에 수렴하는 정도가 약하며 각 변수의 값의 변화도 세대의 진행에 따라서 매우 심하게 나타남을 알 수 있다.

[그림 3]은 유전 알고리즘을 수행하면서 각 세대별로 군집성 분석을 수행한 후 최적값을 구한 결과로 세대가 진행함에 따라서 한가지 값에 빠르게 수렴하며 실재 최적치인 90에도 가깝게 모여 있음을 알 수 있다.

[그림 4]는 두 가지 알고리즘을 식 (8)의 효율성 평가 측도에 의하여 비교한 것으로서 군집분석에 의하여 추정된 최적값이 단순 유전 알고리즘만을 적용한 경우보다 월등히 효과적임을 알 수 있다. 특히 단순 유전 알고리즘에 의한 해는 평균적인 효율성도 떨어질 뿐만이 아니라 60세대가 경과한 후에도 20도 정도의 오차를 갖는 추정치를 선택할 가능성을 보이고 있다. 그러나 유전 알고리즘과 군집성 분석을 동시에 수행한 결과는 40세대 이상에서 7.5도 이상의 오차를 갖는 경우가 나타나지 않고 있다.

[그림 5]는 세대의 진행에 따른 군집의 수로서 세대가 진행됨에 따라서 군집의 수가 초기에 200에서 점점 줄다가 70 세대 부근에서 군집의 수가 40으로 수렴하고 있음을 볼 수 있다. 이는 세대가 진행됨에 따라서 개체가 소수의 국부 피크치를 중심으로 모여 드는 경향이 군집의 수의 감소로 나타나는 것으로 판단할 수 있다.

이러한 결과들을 종합해 볼 때 유전 알고리즘과 군집분석을 적절히 조화함으로써 최적값의 탐색 오차를 크게 줄일 수 있음을 알 수 있다.

5. 결론

본 연구에서는 군집성 분석을 도입하여 성공확률 최적화를 위하여 단순 유전 알고리즘의 정밀도를 보완한 알고리즘을 제안하였으며, 제안된 알고리즘을 테스트 함수에 대하여 적용해 본 결과 단순 유전 알고리즘을 적용한 경우보다 우수한 탐색 능력을 가짐을 알 수 있었다.

본 연구는 특정한 테스트 함수 및 매개 변수에 대한 시뮬레이션 결과로서, 일반적으로 적용되기 위해서는 유전 알고리즘의 여러 매개 변수의 변화나 군집 분석 알고리즘의 변화에 따라서 어떠한 영향을 받는지에 대한 체계적인 추가 연구가 필요하다.

[참고 문헌]

[1] David Beasley, David R. Bull and Ralph R. Martin, *An overview of genetic algorithms : Fundamentals*, Morgan Kaufmann, 1993

[2] David E. Goldberg, *Genetic algorithms in search, optimizations & machine learning*, Addison Wesley Co., 1989.

[3] 허성필, 이동훈, "유전 알고리즘(genetic algorithm)을 이용한 시뮬레이션 최적화 기법", *98 추계공동학술대회 논문집*, 1998.

[4]Richard A. Johnson and Winchern Dean W., *Applied Multivariate Statistical Analysis*, Prentice Hall, 1982.