# 유전자 알고리즘과 군집 분석을 이용한 확률적 시뮬레이션 최적화 기법

# Genetic Algorithm and Clustering Technique for Optimization of Stochastic Simulation

이 동 훈 · 허 성 필
국방과학연구소 · 해군사관학교

요 약

유전자 알고리즘은 전통적인 등반 알고리즘을 이용하여 구하기 어려웠던 최적화 문제를 해결하기 위한 강인한(Robust) 탐색 기법이다. 특히 목적함수가 (1)여러 개의 국부 최대치를 가지는 경우, (2)수학적으로 표현이 불가능하거나 어려운 경우, (3)목적함수에 교란 항(disturbance term)이 섞여 있을 경우도 우수한 탐색 능력을 갖는 것으로 알려져 있다.

본 논문에서는 유전자 알고리즘을 이용하여 나타나는 다양한 해집합을 형성하는 개체군을 군집성 분석(cluster analysis)을 이용하여 군집화하고, 각 군집에 부여된 군집 적합도에 따라서 최적해를 구함으로써 단순 유전자 알고리즘에 의한 최적화보다 훨씬 향상된 탐색 알고리즘을 제안하였다. 반응표면의 형태가 정형화한 테스트 함수의 형태로 나타난다고 가정한 경우에 대하여 몬테 칼로 시뮬레이션을 통하여 본 알고리즘을 적용하여 평가하고 분석하였다.

Keywords : 유전자 알고리즘(Genetic Algorithm), 적합도(fitness), 군집 적합도(cluster fitness), 확률적 시뮬레이션(Stochastic Simulation), Robust Search Algorithm, 군집 분석(Cluster Analysis), Bernoulli Random Variable, Monte Carlo Simulation.

## 1. 서 론

컴퓨터를 이용한 시뮬레이션은 시간적으로 또는 기술적으로 구현하기 어려운 해석적인 방법에 대한 대안으로 활용된다. 일반적으로 적용되는 고전적인 등반 알고리즘(hill climbing algorithm)은 반응표면(response surface)이 단일 피크(peak)를 가지며 미분 가능하다는 가정 하에 이용되나, 시뮬레이션 모델 구축의 초기 단계에는 최적화 하고자 하는 시뮬레이션의 반응면이 단일 피크를 갖는지 여부와 미분가능성 여부를 판단하기 어려우며, 또한 시뮬레이션의 출력이 성공 또는 실패로 나타나는 경우의 반응면은 베르누이 시행(Bernoulli trial)을 따르는 확률변수(random variable)가 되어서 최적화 기법으로서 전통적인 등반 알

고리즘의 적용이 어려워진다.

고전적인 탐색 방법의 대안으로서 유전자 알고리즘은 인공지능적 기법 중의 하나이며 다양한 후보 해를 제공하고 후보 해들간의 정보 교환으로 최적해에 수렴하는 속도를 높이는 강인한(robust) 탐색 알고리즘으로 알려져 있다.[1,2]

본 논문에서는 (1)반응표면의 형태를 예상하기 어려우며 여러 개의 국부 최대치(local maximum)를 가지는 것으로 가정되며, (2)출력의 형태가 베르누이 확률변수로 나타나며, (3)수학적으로 반응표면을 표현할 수 없는 확률적 시뮬레이션(stochastic simulation)을 이용하여 최적해를 찾는 알고리즘에 대하여 다루고자 한다.

최적해를 찾는 문제는 최종적 해의 정확성 뿐 만이 아니라 해에 근접해 가는 속도가 중요한 요소이다. 허성필[3]은 단순 유전자 알고리즘(simple genetic algorithm)을 변형한 방법을 제안하였으며, 매개변수에 따른 탐색 능력의 변화를 몬테 칼로 시뮬레이션을 통하여 분석하여 확률적 시뮬레이션을 이용한 탐색 전략을 제시하였다.

본 논문에서는 변형된 단순 유전자 알고리즘을 통하여 구한 다양한 해집합에 대하여 다시 군집성 분석(Cluster Analysis)을 실시함으로써 실재 해에 수렴하는 속도를 향상시킨 알고리즘을 제안하였으며 몬테 칼로 시뮬레이션을 통하여 알고리즘의 효율성을 평가하였다.

## 2. 확률적 시뮬레이션의 성공확률 최대화를 위한 유전자 알고리즘

제안된 알고리즘은 성공 여부로 출력이 나타나는 시뮬레이션의 성공확률을 최대화시키는 입력조건을 찾기 위한 것으로, 변형된 단순 유전자 알고리즘을 이용하여 세대(generation)를 진행시켜 감에 따라서 개체군(population) 내의 각 개체(individual)의 적합도(fitness)를 향상시켜 나가는 유전자 알고리즘과 세대별 개체군 내의 개체들을 몇 개의 그룹으로 군집화하고 군집 적합도(cluster fitness)를 부여하여 최적의 군집 적합도를 갖는 군집을 최적해로 선택하는 군집 분석(cluster analysis) 알고리즘으로 구성되어 있다.

### (1) 유전자 알고리즘의 개념

유전자 알고리즘은 진화론의 적자생존(survival of fittest)과 자연도태(natural selection) 등의 유전학적 원리를 이용한 탐색 기법으로 발견적 기법으로 분류할 수 있다.

유전자 알고리즘의 기본 개념은 간단하고 매우 쉬우며, 교차 변이(crossover)와 돌연변이(mutation) 등을 수행하여 다음 세대의 탐색 영역을 결정하는 기법이 중요한 역할을 한다.(교차변이나 돌연변이 등의 구체적인 방법은 참고문헌 [2]를 참조)

### (2) 적용 유전자 알고리즘

본 논문에서 적용된 유전자 알고리즘의 절차는 다음과 같다.

① 단계 1(초기화).

세대 k = 0, 랜덤으로 popsize개의 개체를 선택한다.

$$X_{ik} = \left( x_{0ik}, x_{1ik}, \ldots, x_{pik} \right), i = 1,2\ldots, popsize$$

여기서, k번째 세대의 i번 개체는 p개의 속성으로 표현하며, $x_{hik}$ 는 k번째 세대의 $i$번 개체의 h번째 속성의 값을 나타낸다.

② 단계 2(성공확률 추정 시뮬레이션 수행 및 적합도 계산)

각 개체에 대하여 $X_i$ 의 입력 조건으로 n회의 시뮬레이션을 수행하여, 누적성공횟수(cumulative number of successes) $M_{ik}$ 와 누적시행횟수(cumulative number of trials) $N_{ik}$ 를 각각 식 (1)과 (2)와 같이 구하며,

$$\text{누적성공횟수} : M_{ik} = M_{ik-1} + m_{ik} \quad (1)$$

$$\text{누적시행횟수} : N_{ik} = N_{ik-1} + n \quad (2)$$

여기서, $m_{ik}$ 는 k번째 세대의 i번째 개체에 대하여 시뮬레이션을 실시한 n회의 시행 중 성공횟수를 의미한다.

식 (1)과 (2)를 이용하여 식 (3)과 같은 성공확률의 추정치를 구한다.

$$\text{성공확률 추정치} : \hat{p}_{ik} = M_{ik} / N_{ik} \quad (3)$$

단계 3에서 설명될 다음 세대의 선택 방법에 의하면 우수한 적합도를 갖는 일부 개체는 다음 세대로 속성이 그대로 유지된 채로 유전되며, k번째 개체가 이전 세대에서의 개체 속성 값이 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. 이러한 알고리즘은 반복 횟수가 누적된 우수한 개체는 실재 성공확률에서 벗어날 확률이 작아지므로 우연히 도태될 가능성이 작아지며, 우연히 선택된 열등한 개체는 다음 세대에서 사라질 가능성이 커지게 한다.

③ 단계 3(다음 세대(next generation)의 선택 1)

세대 수 k를 1 증가시킨다. 적합도 상위 (1-pcross)*100%의 개체는 다음 세대로 속성 값의 변화 없이 그대로 보낸다. 여기서, pcross는 적용된 교차확률을 의미한다.

④ 단계 4(다음 세대의 선택 2)

pcross*100%개의 개체는 각 개체가 식 (4)와 같은 선택 확률을 갖는 확률바퀴법(roulette selection)에 따라서 선택하고, 선택된 개체들은 2개씩 짝을 지은 후 교차 변이(crossover)를 이점교차(two-point crossover)로 수행하며, 교차변이된 개체들은 각각 pmutation의 확률에 따라서 돌연변이를 수행함으로써 새로운 개체를 발생시킨다. 여기서, pmutation은 돌연변이 확률을 의미한다.

$$S_{ik} = f_{ik} / \sum_{i=1}^{popsize} f_{ik} \quad (4)$$

이 단계에서 생성된 개체에 해당하는 $N_{ik}$ 와 $M_{ik}$ 는 처음 나타나게 되는 개체들이므로 값을 모두 0으로 놓는다.

⑤ 단계 5.

단계 2 - 단계 4를 최대개체수 maxgen만큼 반복 수행 한 후 최종적으로 선택된 개체군을 이용하여 군집성 분석을 실시하여 최적해를 구한다.

### (3) 적합도 함수

1회의 결과가 성공 또는 실패의 형태로 나타나는 시뮬레이션은 성공확률 p인 베르누이 시행으로 볼 수 있으며, N회의 베르누이 시행에서의 성공횟수 M의 분포는 식 (5)와 같으며, 근사적으로는 식 (6)과 같다.

$$M = N\hat{p} \sim Binom(N, p) \quad (5)$$

$$\hat{p} = M/N \overset{app}{\sim} Normal\left(p, \frac{p(1-p)}{N}\right) \quad (6)$$

N이 유한할 경우 $\hat{p}_{ik}$ 는 $\sqrt{\frac{p(1-p)}{N}}$ 의 오차로 $p_{ik}$ 를 추정하게 되며, 따라서 유전적으로 개체를 선택할 경우 낮은 성공확률을 갖는 개체가 채택되거나 높은 성공확률을 갖는 개체가 탈락될 수도 있게 한다.

N이 커지면 $\sqrt{\frac{p(1-p)}{N}}$ 이 작아지므로, N을 크게 하여 $\hat{p}_{ik}$ 가 낮은 분산을 가지도록 함으로써 우수한 개체가 도태되는 확률을 줄일 수 있게 되어 탐색 능력을 향상시킬 수 있다.

실재성공확률 p의 추정치 $\hat{p}$ 의 $100\alpha\%$ 신뢰 구간의 하한치(lower bound)는 근사적으로 식 (7)과 같이 구해지며, 이 값을 각 개체의 적합도로 정의한다.

$$F = \hat{p} - C/\sqrt{4N} \quad (7)$$

$\hat{p} = M/N$ 대신에 F를 이용하는 것은 $\hat{p} = M/N$ 이 작은 값을 갖더라도 N이 클 경우 여러 세대를 거쳐서 유전된 개체이므로 높은 적합도를 주는 것이 바람직하기 때문이다.

여기서, C는 새로운 개체에게 penalty를 주기 위해 설정하는 상수로서 C를 너무 크게 설정할 경우 처음 들어오는 개체는 우수한 개체라도 기존의 열등한 개체를 도태시키고 최종적인 해집합의 하나로 남기가 어려우며, 반대로 C가 너무 작을 경우 우수한 개체가 열등한 개체가 우연히 얻은 높은 적합도 때문에 도태지 않고 남을 확률이 높아 질 수 있다. 따라서 적절한 C를 설정함으로써 효과적인 탐색 전략에 영향을 줄 수 있다.

### (4) 군집성 분석 및 군집 적합도

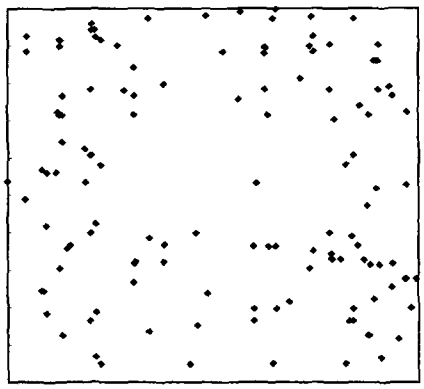
유전자 알고리즘의 특징 중의 하나는 한 개의 최적해를 제시하는 것이 아니고 다양한 복수의 해를 동시에 제공해 주는 것을 장점으로 가지고 있다.

[그림 1]은 유전자 알고리즘을 이용하여 세대의 진행에 따른 해집합의 변화를 나타내는 한 예이다. 세대가 진행되어 감에 따라서 열등한 개체는 소멸되어 가고 대부분의 개체가 몇 개의 그룹으로 모이는 현상을 보이고 있다.

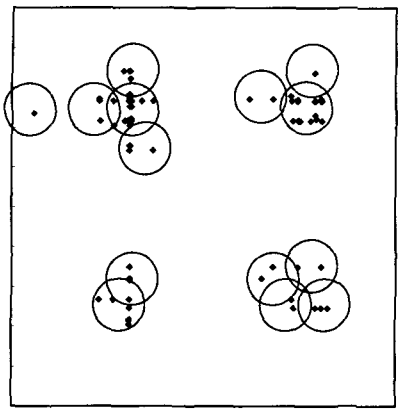
이러한 다양한 해가 모두 국부 피크치라고 보기에는 무리가 있고 상당수가 국부 피크치의 근방에 위치한 개체라고 볼 수 있다.

이럴 경우 이들 국부 피크치 주의의 값 중 하나를 택하여 최적해로 선택하는 것보다 이들의 평균을 이용하는 것이 보다 효과적일 수 있다.

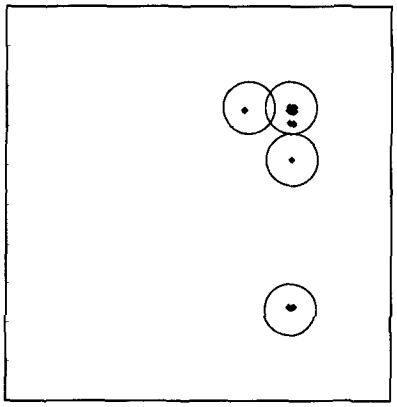
군집 분석(cluster analysis)[4]은 그룹의 수나 구조에 대한 아무런 가정이 없이 유사성(similarity measure)만으로 각 관측치(개체)들을 임의의 그룹에 할당하는 기법으로 하나의 국부 피크치를 중심으로 갖는 개체들을 그룹화하는 알고리즘으로 적용이 가능하다.



(가) 1세대



(나) 10세대



(다) 100세대

[그림 1] 세대의 진행에 따른 유전자 알고리즘에 의한 해집합의 분포의 변화 이러한 유전자 알고리즘의 현상을 이용하기 위한 군집성 분석 및 군집 적합도 부여 절차는 다음과 같다.

① 단계 1.

선정된 개체들을 이용하여 n_cluster개의 군집으로 나눈다. 개체군을 n_cluster개의 군집으로 나누는 절차는 다음과 같다.

- 군집수 n_cluster를 2로 놓는다.
- 각 개체들은 초기 군집으로 나눈다.
- 각 군집의 대표치(centroid) $C_i$를 구한다.

$$C_i = \left( \bar{c}_{i0}, \bar{c}_{i1}, ..., \bar{c}_{ip} \right), i = 1,2,...,n\_cluster$$

$$\bar{c}_{ij} = \sum_{k=1}^{\eta_i} c_{ijk} N_{ik}^* / \sum_{k=1}^{\eta_i} N_{ik}^*,$$

$$i = 1,...,n\_cluster, j = 0,..., p$$

여기서, $c_{ijk}$는 i번째 군집의 k번째 개체의 j번째 변수의 값, $N_{ik}^*$는 i번째 군집의 k번째 군집의 개체의 수, $\eta_i$는 i번째 군집의 개체의 수이다.

- 대표치에 가장 인접한 군집으로 각 개체를 재 할당한다. 인접한 정도는 대표치와 각 개체간의 거리(Euclidean distance)를 계산하여 이용한다.
- 3)과 4)를 반복하여 대표치와 개체간의 거리의 합이 최소가 될 때까지 반복한다.
- 군집 대표치로부터의 거리가 R이상인 개체가 존재할 경우 n_cluster를 하나 증가시키고 2)에서 5)를 반복 수행한다. 최대 거리가 R이하일 경우 군집 분석을 중단한다.

② 단계 2.

단계 1에서 구한 군집의 평균을 이용하여 군집화된 개체를 구하며 식 (8)과 같이 군집 적합도를 구한다.

$$F_{cluster}(i) = \hat{p}_{cluster}(i) - 3/\sqrt{4N_{cluster}(i)} \quad (8)$$

여기서,

$$\hat{p}_{cluster}(i) = M_{cluster}(i) / N_{cluster}(i)$$

$$M_{cluster}(i)=\sum_{k=1}^{\eta_i} M_{ik}^{*},\ N_{cluster}(i)=\sum_{k=1}^{\eta_i} N_{ik}^{*}$$

③ 단계 3.

최대의 군집 적합도를 갖는 군집화된 개체를 최적해로 선정한다.

## 3. 테스트 함수(test function)에 대한 알고리즘 적용

제안된 알고리즘을 이용하여 식 (9)와 같이 주어진 테스트 함수의 g=4인 경우에 대하여 얼마나 효과적으로 최적해를 탐색할 수 있는지를 몬테 칼로 시뮬레이션을 통하여 평가하였다.

### (1) 테스트 함수

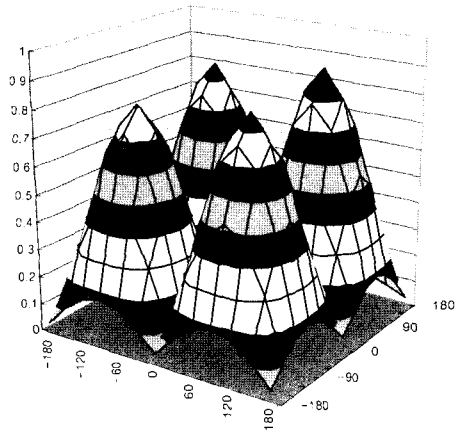
반응표면의 표면이 여러 개의 국부 최대치를 갖거나 확률 변수의 형태를 갖는 경우의 탐색 알고리즘의 효과도 비교를 위한 테스트 함수는 여러 가지가 알려져 있으나[2] 여기서는 식 (9)과 같은 테스트 함수를 정의하고 유전자 알고리즘의 효과도를 알아보았다.

$$p[x]=0.9\times\prod_{i=0}^{g} f(x_i) \quad \text{----------(9)}$$

여기서,

$$f(ix)=\begin{pmatrix} -09\times\sin(\xi),\ \text{if}\ -180<\xi<0 \\ \sin(\xi)\ \text{if}\ 0<\xi<180 \end{pmatrix}$$

[그림 2]는 g=1인 경우의 테스트 함수의 형태를 보여주고 있다. 변수의 값이 모두 90을 가질 경우 최적값을 가지며, 한 개 이상의 변수가 -90일 경우에 최적값보다 다소 작은 출력을 내는 국부 피크치를 갖는 함수이다. 이 테스트 함수는 $2^{g+1}$개의 국부 최대치를 가지게 된다.




〔그림 2〕 테스트 함수의 형태(g=1)

### (2) 알고리즘에 대한 몬테 칼로(Monte Carlo) 시험계획

유전자 알고리즘의 최적해를 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 최적해로 선정하였다.

〔표 1〕 적용 유전자 알고리즘의 매개변수

| 매개 변수 | 값 |
|---|---|
| 개체군의 크기 | 200 |
| 최대세대수 | 100 |
| 개체당 시행의 수 | 6 |
| 교차 확률 | 0.6 |
| 돌연변이 확률 | 0.003 |
| C | 0.5 |

한 세대의 시뮬레이션 횟수와 관계되는 개체군의 크기(population size)와 개체 당 시뮬레이션 횟수는 각각 200과 6이다. 따라서 한 세대에 1,200회의 시뮬레이션이 필요하며, 100 세대까지 진행될 경우 120,000회의 시뮬레이션을 수행하게 된다.

교차확률(crossover probability), 돌연변이 확률(mutation probability)과 C값은 허성필[3]의 분석 결과에 따라서 비교적 효율적인 값으로 선정하였으며, 군집분석에서 적용될 R 값은 10으로 적용하여 모든 개체가 군집의 centroid로부터 10도 이내에 있을 경우 군집의 수를 증가시키지 않도록 하였다.

알고리즘의 효율성을 평가하는 측도(measure)는 식(10)과 같은 개념의 알고 있는 최적해와 추정된 최적해 간의 거리를 이용한다.

$$D=\sqrt{\sum_{i=0}^{g}(\mu_i-\hat{\mu}_i)^2} \quad \text{----------(10)}$$

여기서, $\mu_i$와$\hat{\mu}_i$는 각각 알고 있는 최적해와 알고리즘에 의해서 추정된 최적해이다.

### (3) 몬테 칼로 시뮬레이션 결과 분석

[그림 3]은 변형된 단순 유전자 알고리즘을 적용할 경우에 세대의 진행에 따라서 최적해를 찾아 나가는 과정을 보여주는 것으로, 세대의 진행에 따라서 일정한 값에 수렴하는 정도가 약하며 세대가 상당히 진행되어도 각 변수의 값의 변화가 여전히 매우 심하게 나타나며 돌출적인 값들이 계속적으로 나타남을 알 수 있다.

[그림 4]는 유전자 알고리즘을 수행하면서 각 세대별로 군집성 분석을 수행한 후 최적값을 구한 결과로 세대가 진행함에 따라서 한가지 값에 빠르게 수렴하며 실재 최적치인 90에도 가깝게 모여들고 있음을 알 수 있다.

[그림 5]는 세대별 최적해 추정치의 추세를 5개의 변수 중 특정 변수 $x_3$에 대하여 그린 것으로, 비록 실재해 90에 수렴은 못 하고 있지만 단순 유전자 알고리즘만을 적용한 경우보다 군집분석을 가미한 기법이 변동폭이 작음을 알 수 있다.

[그림 6]은 두 가지 알고리즘을 식 (8)의 효율성 평가 측도에 의하여 비교한 것으로서 군집분석에 의하여 추정된 최적해가 단순 유전자 알고리즘만을 적용한 경우보다 월등히 효과적임을 알 수 있다.

특히 단순 유전자 알고리즘에 의한 최적해는 평균적인 효율성도 떨어질 뿐만이 아니라 60세대가 경과한 후에도 20도 정도의 오차를 갖는 추정치를 선택할 가능성을 보이고 있다. 그러나 유전자 알고리즘과 군집성 분석을 동시에 수행한 결과는 40세대 이상에서 7.5도 이상의 오차를 갖는 경우가 나타나지 않고 있다.

[그림 7]은 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 모여드는 경향이 군집의 수의 감소로 나타나는 것으로 판단할 수 있다.

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

## 5. 결론

본 연구에서는 여러 개의 국부 피크치를 가지며 출력이 베르누이 확률변수의 형태로 나타나는 확률적 시뮬레이션의 최적화 문제를 해결하기 위하여, 단순 유전자 알고리즘에 군집성 분석 알고리즘을 도입하여

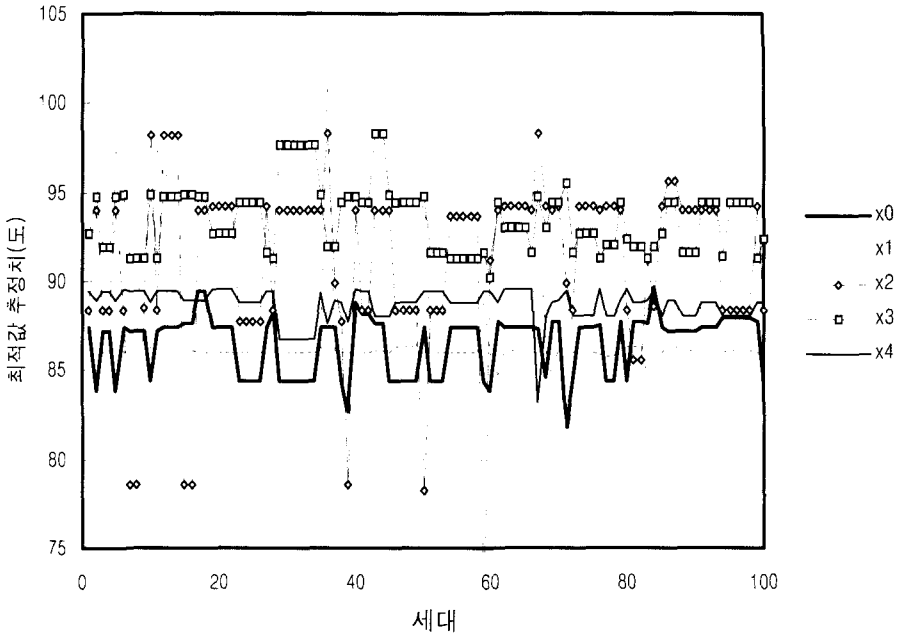
탐색 속도를 향상시킨 알고리즘을 제안하였다.

제안된 알고리즘을 정형화된 테스트 함수에 대하여 적용해 본 결과 단순 유전자 알고리즘을 적용한 경우보다 우수한 탐색 능력을 가짐을 알 수 있었으며, 확률적 시뮬레이션 문제에 있어서의 시행 횟수를 줄이는 데에 크게 기여할 수 있음을 확인하였다.

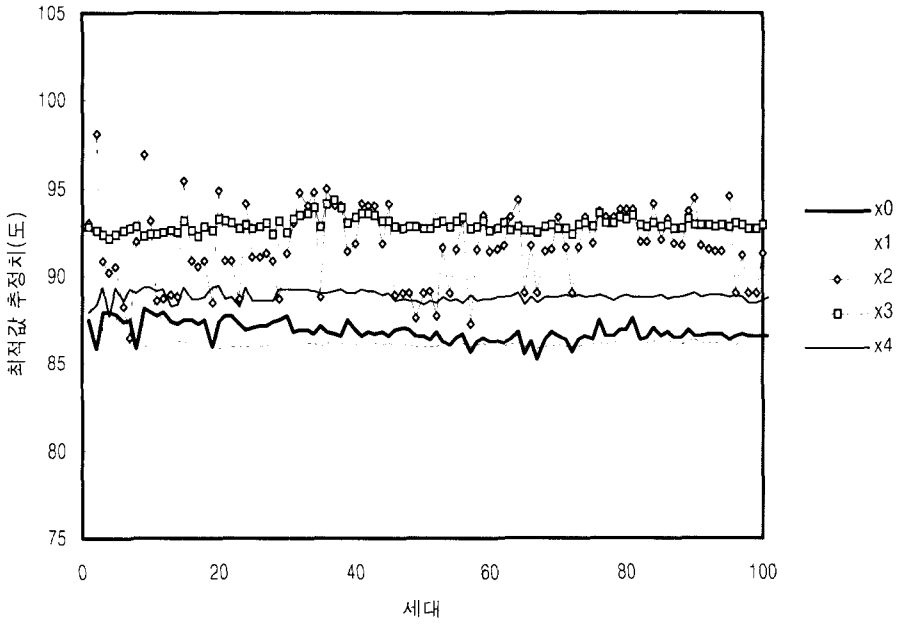
본 연구의 결과는 정형화된 하나의 테스트 함수에 대하여 적용한 것이며 일반적으로 적용되기 위한 추가적인 연구가 필요하며, 시뮬레이션의 출력의 형태가 베르누이 시행의 형태 외의 다른 형태의 확률변수로 나타날 경우에 대한 연구도 병행되어야 하겠다.

## 참 고 문 헌

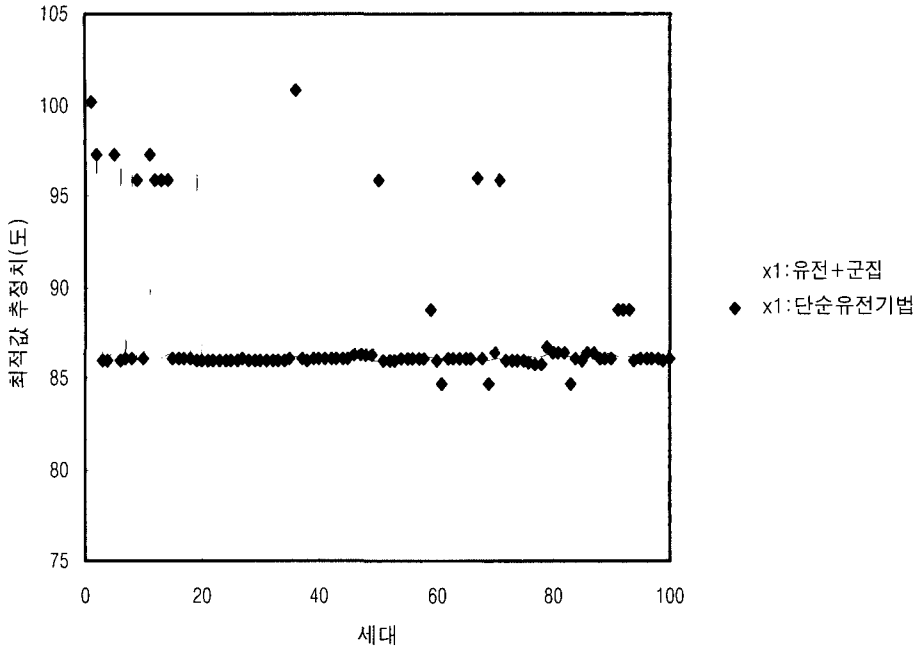

1. David Beasley, David R. Bull and Ralph R. Martin, An overview of genetic algorithms : Fundamentals, Morgan Kaufmann, 1993
2. David E. Goldberg, Genetic algorithms in search, optimi-zations & machine learning, Addison Wesley Co., 1989.
3. 허성필, 이동훈, 유전 알고리즘(genetic algorithm)을 이용한 시뮬레이션 최적화 기법, 98추계공동학술대회 논문집, 1998.
4. Richard A. Johnson and Winchern Dean W., Applied Multivariate Statistical Analysis, Prentice Hall, 1982.



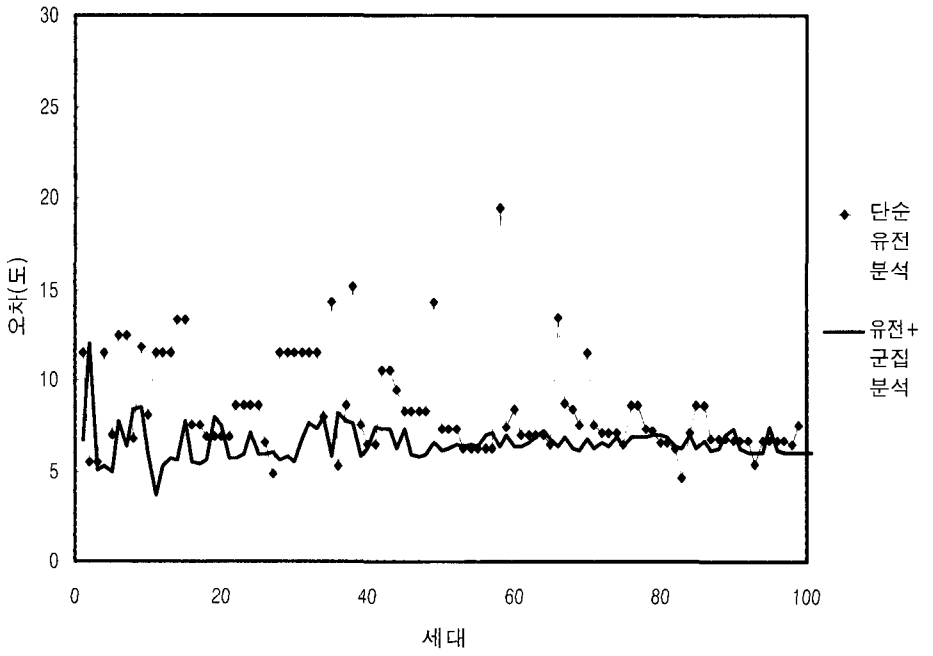

〔그림 3〕 단순 유전자 알고리즘을 이용한 경우의 세대에 따른 최적값 추정치의 추세



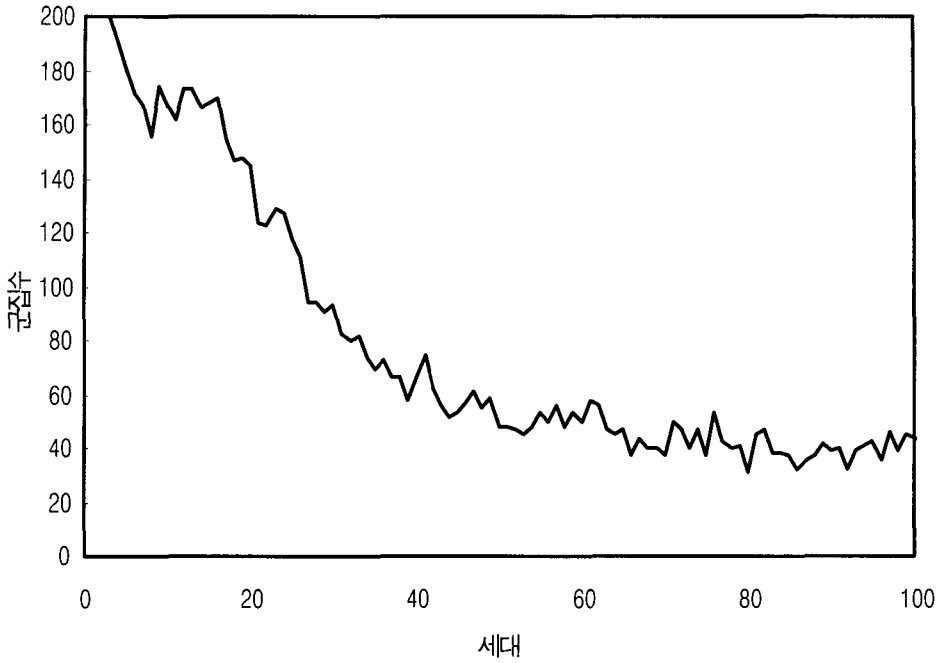

〔그림 4〕 유전자 알고리즘과 군집분석을 이용한 경우의 세대에 따른 최적값 추정치의 추세




〔그림 5〕 세대의 진행에 따른 특정변수에 대한 최적값 추정결과 비교 ($x_3$)




〔그림 6〕 세대에 따른 오차의 추세비교




〔그림 7〕 세대에 따른 군집의 수