개체 클러스터링을 이용한 효율적인 국소 평가 유전자 알고리즘

김희수 조성배
연세대학교 컴퓨터과학과
{madoka,sbcho}@candy.yonsei.ac.kr

An Efficient Genetic Algorithm with Partial Evaluation by Clustering

Hee-Su Kim and Sung-Bae Cho
Computer Science Department, Yonsei University

요 약

유전자 알고리즘을 적용하는 문제의 경우 일반적으로 집단의 크기를 가능한 한 크게 유지시킴으로써 최적의 해가 찾아지도록 한다. 그러나 개체 평가 비용이 상대적으로 큰 몇몇 특정한 문제의 경우 집단의 크기가 커지면 심각한 문제가 되기도 한다. 이러한 이유로 본 논문에서는 클러스터링 기법을 이용한 국소 평가 유전자 알고리즘을 제안하였다. 이 방법은 집단을 몇 개의 클러스터로 나누고 각각의 대표 개체를 평가한 후 나머지 개체들의 적합도 값은 간접적인 계산에 의해 얻어내는 방법으로, 적은 수의 평가만으로도 상대적으로 큰 집단을 유지시키는 효과를 얻을 수 있다. 일반적인 유전자 알고리즘과의 성능 비교를 통해 제안된 알고리즘이 효율적이었음을 알 수 있었다.

1. 서론

유전자 알고리즘은 적자생존, 유전자 교차 등의 자연 진화 메카니즘에 기반한 문제해결 알고리즘으로 기계학습, 최적화 및 분류 작업에 매우 효율적인 해결 방법을 제시해 준다. 이 방법은 문제에 맞게 인코딩된 다수의 개체들을 적합도 함수에 의해 평가하여 우수한 개체들을 선택한 후 교차, 돌연변이 연산을 적용하는 과정을 반복함으로써 집단을 최적화시키는 방법이다. 이러한 방법은 상대적으로 넓은 탐색공간에 분포한 개체 집단을 진화시키므로 가능한 한 집단의 크기를 크게 유지하는 것이 바람직하다. 그러나 개체 평가에 드는 비용이 상대적으로 큰 일부 특정한 문제의 경우 집단을 크게 하기가 곤란하므로 일반적인 유전자 알고리즘보다 작은 집단을 통해 탐색을 수행하여야 하며, 그 결과 탐색과정에서 해의 쏠림 현상과 같은 문제가 발생하게 된다.

이러한 문제의 대표적인 예가 대화형 유전자 알고리즘 및 그 응용 분야들이다. 일반적인 유전자 알고리즘이 예술이나 감성적인 문제 등에는 적합도 함수를 도출하기 어렵다는 점 때문에 적용이 곤란한 반면 대화형 유전자 알고리즘은 사용자의 평가를 직접 적합도 함수로 사용함으로써 사용자의 선호도나 감성을 진화에 반영시킬 수 있으며, 따라서 음악이나 디자인과 같은 영역의 문제 해결에도 적용이 가능하다. 그러나 이러한 접근 방법은 사용자가 직접 각 개체에 대한 평가를 해야 한다는 점 때문에 집단의 크기가 사용자가 제어할 수 있는 정도로만 제한되어지며, 따라서 원하는 결과를 얻기 어렵다.

본 논문에서는 이러한 문제를 해결하기 위해 클러스터링 기법을 이용하여 작은 탐색집단을 가지고도 큰 탐색집단을 이용할 때와 거의 같은 성능을 보이는 국소 평가 유전자 알고리즘을 제안하고자 한다.

2. 클러스터링 기법

클러스터링은 전체 집단을 각각의 특징을 가지는 몇 개의 부집단으로 분할하는 기법이다. 클러스터링 기법은 크게 계층적 클러스터링, 분할 클러스터링, 중복 클러스터링으로 나뉜다. 계층적 클러스터링은 부집단들이 더 작은 부집단으로 이루어진 하부 구조를 가지도록 하는 방법이며, 다시 bottom-up 방식의 agglomerative 알고리즘과 top-down 방식의 divisive 알고리즘으로 나뉜다. 분할 클러스터링은 부집단간에 계층 구조가 없으면서 각각의 부집단들이 서로에게 닫혀 있도록 하는 기법이며, hard c-means 알고리즘, k-means 알고리즘 등이 이에 속한다. 중복 클러스터링은 부집단간에 계층 구조가 없고 각 부집단이 부분적으로 중복될 수 있으며, fuzzy c-means 알고리즘, b-clump 알고리즘 등이 있다.

클러스터링 기법은 최근 각광받는 화상 처리를 비롯하여 여러 가지 분야에서 다양하게 응용되고 있다.

3. 국소 평가 유전자 알고리즘

본 논문에서는 클러스터링 방법을 이용하여 큰 탐색집단의 일부분만을 직접 평가하면서 전체를 평가하는

경우와 거의 동일한 수준의 성능을 보이는 국소 평가 유전자 알고리즘을 제안하고자 한다. 이러한 방법은 적합도 평가에 드는 비용이 상대적으로 큰 문제를 유전자 알고리즘으로 해결하는 데 매우 유용하다. 제안된 알고리즘은 다음과 같다.

```
Procedure Cluster_GA(){
        initialize ();
                While not end condition do
                SimpleGA();
                Clustering();
                Evaluation();
        End while
}
Procedure SimpleGA(){
        Select();
        Crossover();
        Mutate();
}
Procedure Clustering(){
        /* Performs clustering which divides
        population into k cluster */
}
Procedure Evaluation(){
        Pick_Representatives();
        Evaluate_Representatives();
        Indirect_Evaluation();
}
Procedure Indirect_Evaluation(){
        /* Evaluates non-representative
        individuals indirectly from the
        fitness value of representatives */
}
```

우선 일반적인 유전자 알고리즘에서처럼 비교적 크게 집단의 크기를 설정하여 초기화한다. 이 집단을 대상으로 클러스터링 기법을 이용하여 비슷한 특징의 개체들을 모아 k개의 부집단들을 만든 후, 부집단 내의 대표 개체들의 적합도만을 평가하도록 한다. 집단 내의 다른 개체들에 대한 적합도는 각 개체가 속한 부집단 내의 대표개체의 적합도 값으로부터 간접적으로 계산하여 할당한다.

부집단을 구성하는 클러스터링 기법으로 본 논문에서는 가장 간단하고 직관적인 알고리즘 중 하나인 k-means 알고리즘을 사용하였다. k-means 알고리즘의 개략적인 적용과정은 다음과 같다.

```
1. Form k clusters with first k samples
2. For each of the remaining n-k samples{
        • Put the sample into the cluster
          identified with nearest centroid
        • Recompute the centroid of altered
          cluster
}
3. For each of all n samples{
        • Put the sample into the cluster
          identified with nearest centroid
}
```

한편, 각 개체의 적합도를 간접적으로 산출하는 방법으로는 아래와 같이 Euclidean method를 사용하여 대표값으로부터의 거리를 계산한 후 이에 반비례하게 적합도를 부여하는 방법을 사용하였다.

$$dist(x)=\sqrt{\sum_{i=1}^{n}(c_i-x_i)^2}$$

4. 실험 및 결과 분석

제안된 알고리즘의 성능 평가를 위해 9가지 성능 평가 함수를 이용하여 국소 평가 알고리즘을 포함한 3가지 유전자 알고리즘의 성능을 비교하였다. 실험에 사용된 알고리즘들은 표 1에, 실험에 사용된 환경변수 값들은 표 2에 정리하였다.

표 1 실험에 사용된 유전자 알고리즘

pop100	집단의 크기가 100인 일반 유전자 알고리즘
clu10	집단의 크기가 100이며, 이를 10개의 부집단으로 나누어 10번의 직접 평가만을 수행하는 국소 평가 유전자 알고리즘
pop10	집단의 크기가 10인 일반 유전자 알고리즘

표 2 환경 변수

	pop100	clu10	pop10
집단의 크기	100	100	10
평가 회수	100	10	10
부집단의 수	-	10	-
교차 확률	0.9		
돌연변이 확률	0.001		
세대 수	200		

실험에 사용된 성능 평가 함수들은 다음과 같다.

<Function 1> De Jong function 1

$f(x)=\sum_{i=1}^{n}x_i^2$, n=3, $-5.12\le x_i\le 5.12$

<Function 2> De Jong function 2

$f(x)=100(x_1^2-x_2)^2+(1-x_1)^2$, $-2.048\le x_i\le 2.048$

<Function 3> De Jong function 3

$f(x)=\sum_{i=1}^{n}\mathrm{int}(x_i)$, n=5, $-5.12\le x_i\le 5.12$

<Function 4> De Jong function 4

$$f(x)=\sum_{i=1}^{n}(ix_i^4+Gauss(0,1)),\quad n=30,\ -1.28\le x_i\le 1.28$$

<Function 5> De Jong function 5

$$f(x)=0.002+\sum_{j=1}^{25}\frac{1}{j+\sum_{i=1}^{2}(x_i-a_{ij})^6},\quad -65.536\le x_i\le 65.536,$$

$$(a_{ij})=\begin{pmatrix}-32 & -16 & 0 & 16 & 32 & -32 & \Lambda & -16 & 0 & 16 & 32\\ -32 & -32 & -32 & -32 & -32 & -16 & \Lambda & 16 & 32 & 32 & 32\end{pmatrix}$$

<Function 6> Rastrigin function

$$f(x)=3.0n+\sum_{i=1}^{n}x_i^2-3.0\cos(2\pi x_i),\quad n=20,\ -5.12\le x_i\le 5.12$$

<Function 7> Schwefel function

$$f(x)=418.9829n+\sum_{i=1}^{n}x_i\sin\left(\sqrt{|x_i|}\right),\quad n=10,\ -500.0\le x_i\le 500.0$$

<Function 8> Griewangk function

$$f(x)=1+\sum_{i=1}^{n}\frac{x_i^2}{4000}-\prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right),\quad n=10,\ -600.0\le x_i\le 600.0$$

<Function 9> Ackley function

$$f(x)=20+e-20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\right)-\exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right),$$

$n=30,\ -30.0\le x_i\le 30.0$

그림 1은 사용된 성능 평가 함수의 하나인 Schwefel function의 Landscape이다. 이 함수는 전역해와 멀리 떨어진 곳에 두 번째 근사해가 존재하기 때문에 최적화 알고리즘으로 하여금 이 두 번째 봉우리에 고착되도록 유도한다. 그림 2는 이를 통해 평가한 세 알고리즘의 성능 비교이며, pop100과 clu10이 비교적 빨리 전역해로 수렴하는 데 비해 pop10은 국소해에 쉽게 고착됨을 알 수 있다. 나머지 성능 평가 함수에서도 이와 비슷한 결과를 얻을 수 있었으며, 이로부터 제안된 알고리즘은 10번의 적합도 평가만으로도 100번의 평가를 수행하는 유전자 알고리즘과 대동소이한 성능을 가진다고 할 수 있다. 반면, 10번의 평가를 수행하는 일반 유전자 알고리즘은 이 둘에 훨씬 못 미치는 성능을 보여주었다.

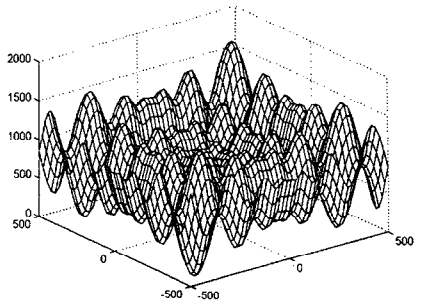

그림 1 Schwefel function의 3-D landscape

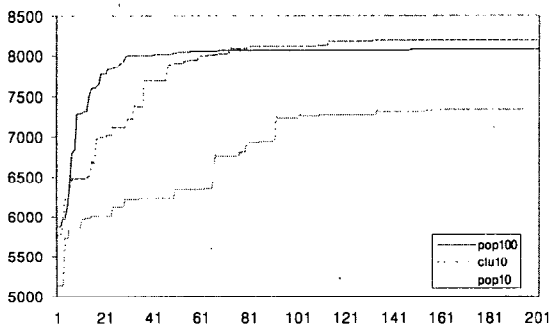

그림 2 Schwefel function에 의한 성능 평가

5. 결론 및 향후 연구

본 연구에서는 클러스터링을 통해 적은 수의 평가만으로도 상대적으로 큰 집단을 유지하는 것과 거의 동일한 성능을 보여주는 국소 평가 유전자 알고리즘을 제안하였다. 집단은 몇 개의 부집단으로 나누어지며, 각 부집단의 대표값들에 대한 평가가 이루어진 후 나머지 개체들은 이 값으로부터 계산에 의한 간접적인 적합도 값을 부여받는다. 다수의 성능평가 함수를 통한 비교에서 제안된 알고리즘이 큰 집단을 유지하는 알고리즘과 비슷한 수준의 성능을 가짐을 증명하였다.

이러한 접근 방법은 개체 평가에 드는 비용이 상대적으로 높은 문제의 해결에 유용하며, 특히 사용자가 직접 개체를 평가하기 때문에 집단의 크기가 제한되는 대화형 유전자 알고리즘에 매우 효과적이다. 이 경우 사용자가 실제로 평가하는 개체 수는 늘리지 않으면서 일반 유전자 알고리즘처럼 큰 집단을 유지하는 효과를 얻을 수 있다.

본 연구에서는 클러스터링 방법으로 k-means 알고리즘만을 사용하였으나 향후 연구에서는 SOM이나 fuzzy c-means 알고리즘과 같은 다른 클러스터링 기법으로 이를 대체하여 그 성능을 비교해보려 하며, 궁극적으로는 제안된 국소 평가 기법을 대화형 유전자 알고리즘과 결합하여 성능 향상을 꾀하고자 한다.

참고 문헌

[1] Fukunaka, K., *Introduction to Statistical Pattern Recognition*, Academic Press, New York 1990.

[2] Goldberg, D.E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Co. Inc., 1989.

[3] Gose, E., Johnsonbaugh, R. and Jost, S., *Pattern Recognition and Image Analysis*, Prentice Hall PTR, 1996.

[4] Takagi, H., "Interactive evolutionary computation: Cooperation of computational intelligence and human KANSEI," *Proc. of Int'l Conf. on Soft Computing*, pp. 41~50, 1998.