# 유전 알고리즘을 이용한 FCM 알고리즘의 초기 군집 중심 선택

오종상, 정순원, 박 귀태
서울 성북구 안암동 5가 1 고려대학교 전기공학과
ojs@eeserver.korea.ac.kr

# A initial cluster center selection in FCM algorithm using the Genetic Algorithms

Jong-Sang Oh, Soon-Won Jung, Gwi-Tae Park
Dept. of Electrical Eng. Korea Univ. 5-1 Anam-dong, Sungbuk-gu, Seoul
ojs@eeserver.korea.ac.kr

Abstract

This paper proposes a scheme of initial cluster center selection in FCM algorithm using the genetic algorithms. The FCM algorithm often fails in the search for global optimum because it is local search techniques that search for the optimum by using hill-climbing procedures. To solve this problem, we search for a hypersphere encircling each clusters whose parameters are estimated by the genetic algorithms. Then instead of a randomized initialization for fuzzy partition matrix in FCM algorithm, we initialize each cluster center by the center of a searched hypersphere.

Our experimental results show that the proposed initializing scheme has higher probabilities of finding the global or near global optimal solutions than the traditional FCM algorithm.

## 1. 서론

패턴(pattern) 공간상에 다수의 패턴 데이터가 서로 가깝게 무리를 이룰 때 이 무리를 이루는 패턴 데이터의 집합을 군집(cluster)이라 한다. 그리고, 각 패턴간의 유사성이나 근접성을 이용하여 주어진 패턴을 무리지우는 과정을 군집화(clustering)라고 말한다. 목적 함수 값을 최소로 하는 대표적인 퍼지 군집화 알고리즘으로는 FCM(Fuzzy C-Means) 알고리즘이 있다.[1] 그러나, 이 방법은 언덕 등반 방법(hill-climbing procedure)을 통해 최적값을 탐색하는 지역 탐색 방법(local search technique)이므로 전역 최적값 탐색을 보장하지 못하는 단점이 있다. 초기 군집 중심을 어떻게 설정하느냐에 따라 전역 최적값 탐색 여부가 결정되고, 이는 데이터의 차수가 커지고 분포가 복잡해질 때 적절한 초기 군집 중심 설정은 거의 불가능한 일이라고 할 수 있다. 따라서, 보통 FCM 알고리즘 수행시 퍼지 분할 행렬을 랜덤하게 초기화시키는 방법이 사용된다.

파라미터 최적화 기법인 유전 알고리즘은 자연계의 진화 과정을 흉내낸 구조적인 랜덤 탐색 알고리즘의 하나로서, 최적화가 필요한 여러 분야에 적용되어 좋은 결과를 보여 주고 있다.[2]

본 논문에서는 패턴 공간상에서 각 군집을 둘러싸는 최소 부피를 갖는 초구체(hypersphere)의 중심 즉 군집 중심을 유전 알고리즘으로 탐색해 그 값을 FCM 알고리즘의 초기 군집 중심으로 초기화하는 방법을 제안한다. 패턴 공간상에서 각 군집을 둘러싸는 최소 부피를 갖는 평형 육면체나 회전 타원체를 유전 알고리즘을 통해 구하는 방법은 [3]에서 제안된 바 있으며, 여기서는 이 방법을 패턴 인식 문제에 있어, 결정 경계(decision boundary)를 결정하는 문제에 사용하였다. 이러한 방법에 의한 초기화로 제안된 방법이 퍼지 분할 행렬을 랜덤하게 초기화시키는 기존의 FCM 알고리즘 방법에 비해 지역 최적값에 빠지는 확률을 줄일 수 있음을 실험을 통해 알 수 있었다.

## 2. 유전 알고리즘

### 2.1 단순 유전 알고리즘

유전 알고리즘은 유전적인 계승과 다아윈적 생존 경쟁이라는 자연계의 현상을 모델링한 통계 확률적인 탐색 방법으로 1975년 미국의 John Holland 교수에 의해 제안된 알고리즘이다.

단순 유전 알고리즘(Simple Genetic Algorithms)은 탐색의 대상이 되는 문제에 대한 해를 고정된 크기의 염색체라 불리는 개체로 표현하는 것으로 시작한다. 하나 이상의 개체들이 모여 전체 개체 집단을 이루며, 전 세대를 통하여 고정된 크기의 집단의 수를 유지하게 된다. 그리고, 새로운 세대를 생산하기 위하여 전체 개체 집단의 해는 적합도 함수(fitness function)라는 평가 함수에 의해 각 개체가 평가된다. 진화 과정은 선택 연산자, 교배 연산자 그리고 돌연 변이 연산자의 유전 연산자를 사용한다. 전 진화 과정을 통하여 더 나은 적합도를 가지는 해, 즉 평균 적합도보다 큰 해는 선택 연산에서 지수적으로 더 높은 확률로 선택되게 된다. 점차 세대가 진화해감에 따라 각 개체들은 탐색 공간에 대한 정보를 축적하게 되고 최적이나 최적 근방의 해로 결국에는 수렴하게 된다. 유전 알고리즘은 고유한 병렬 탐색을 행하며 일반적으로 국부 최적값에 잘 빠지지 않는다.

### 2.2 적응적 유전 알고리즘

단순 유전 알고리즘은 어떤 특정한 경우에는 최적해를 찾지 못하고, 국부 최적값에 조기에 수렴할 가능성이 있다. 이러한 문제를 해결하고자 M. Srinivas와 L. M. Patnaik는 교배와 돌연변이 확률을 각각의 개체의 적합도와 개체 집단 전체의 적합도 정보를 이용하여 적응력을 갖는 교배 확률 ($P_c$)과 돌연변이 확률 ($P_m$)로 유전 알고리즘이 국소 최소값으로 수렴하는 문제를 상당히 완화하는 적응적 유전 알고리즘(Adaptive Genetic Algorithms)을 제안하였다.[4]

수렴 성능에 영향을 미치는 두 확률 변수를 적응적으로 변화시켜 적합도가 큰 개체는 교배와 돌연 변이 확률을 작은 값으로 주어 그 개체를 다음 세대로 보존한다. 그러한 반면에 평균 이하의 적합도를 갖는 개체는 새로운 해를 만들도록 큰 확률값으로 주어 파괴시킨다. 최대 적합도 값에서 평균 적합도 값의 차이($f_{\max} - \overline{f}$)는 유전 알고리즘의 수렴성을 판단하는 기준이 되며 이를 이용하여 각 적합도 값에 대한 적응적 확률을 계산한다.

$$P_c = \begin{cases} k_1(f_{\max} - f')/(f_{\max} - \overline{f}) & , f' \geq \overline{f} \\ k_3 & , f' < \overline{f} \end{cases}$$

$$P_m = \begin{cases} k_2(f_{\max} - f)/(f_{\max} - \overline{f}) & , f \geq \overline{f} \\ k_4 & , f < \overline{f} \end{cases}$$

$$where \quad 0 < k_1,\ k_2,\ k_3,\ k_4 \leq 1.0 \quad (1)$$

## 3. FCM 알고리즘과 유전 알고리즘을 이용한 초기 군집 선택 방법

### 3.1 FCM 알고리즘

군집의 중심점과 각 데이터 사이의 거리로 하는 목적함수를 정하고, 이를 최소화하는 알고리즘이 FCM 알고리즘이다. 이 알고리즘을 간단히 설명한다.

$p$ 차원의 특징 패턴 공간을 갖는 $n$ 개의 데이터가 있다고 하자. 이는 식(2)와 같이 나타낼 수 있다.

$$\boldsymbol{X} = \{ \boldsymbol{x}_j ; j = 1, 2, \cdots, n \}, \quad \boldsymbol{x}_j \in R^p \quad (2)$$

분할하고자 하는 군집의 개수 $c (2 \leq c \leq n)$ 이며, 각 군집의 중심은 $c \times p$ 차의 행렬 $\boldsymbol{V}$이며, $c \times n$ 차의 퍼지 분할 행렬 $\boldsymbol{U}$는 식(3~5)와 같이 정의한다.

$i$) 모든 $i$, $j$에 대하여 $\quad \mu_{ij} \in [\ 0,\ 1\ ]$ (3)

$ii$) 모든 $j$에 대하여 $\quad 0 < \sum_j \mu_{ij} < n$ (4)

$iii$) 모든 $i$에 대해여 $\quad \sum_i \mu_{ij} = 1$ (5)

최적의 퍼지 분할 행렬 $\boldsymbol{U}$를 구하기 위한 목적 함수는 식(6)과 같이 정의된다.

$$J_M(U, V) = \sum_{j=1}^{n} \sum_{i=1}^{c} (\mu_{ij})^m (d_{ij})^2 = \sum_{j=1}^{n} \sum_{i=1}^{c} (\mu_{ij})^m \| \boldsymbol{x}_j - \boldsymbol{V}_i \|^2 \quad 1 \leq m \leq \infty \quad (6)$$

$d_{ij}$는 $i$번째 군집 중심과 $j$번째 데이터간의 보통 유클리디안 거리이며, 이는 각 데이터와 군집 중심과의 유사도를 나타낸다. $m$은 퍼지화 상수 또는 지수 가중치(exponential weight)로 불린다.

목적 함수 식(6)을 최소로 하는 퍼지 분할 행렬 $\boldsymbol{U}$, 군집의 원형(prototype) $\boldsymbol{V}$는 식(7)과 식(8)로 계산된다.

$$\boldsymbol{V}_{il} = \frac{\sum_{j=1}^{n} (\mu_{ij})^m \boldsymbol{x}_{jl}}{\sum_{j=1}^{n} (\mu_{ij})^m} \quad (7)$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \quad \text{if} \quad d_{ij} \neq 0 \quad (8)$$

### 3.2 유전 알고리즘을 이용한 FCM 알고리즘의 초기 군집의 선택 방법

본 논문에서 제안하는 방법은 FCM 알고리즘에서 퍼지 분할 행렬 $\boldsymbol{U}$의 계수를 초기화하는 대신에 각 군집을 최소의 부피로 둘러싸는 초구체를 탐색하여 초구체의 중심을 군집의 중심 값 $\boldsymbol{V}$로 초기화하는 것이다. 초구체의 중심을 탐색함에 있어 유전 알고리즘이라는 전역 최적화 알고리즘에 의해서 각 파라미터의 추정이 가능하다. 단순한 탐색 알고리즘에 의해서는 해공간이 너무 커서 불가능하고 기울기 강하법에 의해서는 국부 최적값에 빠질 확률이 높다.

유전 알고리즘을 적용하기 위해서는 첫번째로 각 파라미터를 각 개체로 표현해야 한다. 패턴 공간상의 $p$ 차원의 데이터 $\boldsymbol{X} = \{ \boldsymbol{x}_j ; j = 1, 2, \cdots, n \}$, $\boldsymbol{x}_j \in R^p$가 있다. 그리고 미리 정한 군집의 개수는 $c$ 개라고 하자. $p$ 차원 초구체의 방정식은 식(9)와 같다.

$$\sum_{k=1}^{p} ( x_k - C_k)^2 = R^2 \quad (9)$$

유전 알고리즘에서 추정해야 될 파라미터는 초구체의 중심 $C=\{ C_{ik}; i=1,2,\cdots,c; k=1,2,\cdots,p\}$와 초구체의 반지름 $R=\{ R_i; i=1,2,\cdots,c\}$이다. 그래서, 탐색할 파라미터의 수는 $c(p+1)$개이다. 각 개체의 형태는 그림 1과 같이 표현된다.

| $C_{11}$ | $C_{12}$ | ... | $C_{1p}$ | $C_{21}$ | ... | $C_{c1}$ | ... | $C_{cp}$ | $R_1$ | ... | $R_c$ |
|---|---|---|---|---|---|---|---|---|---|---|---|

그림 1 각 개체의 형태

각 유전 인자의 길이를 8비트(bit) 스트링으로 한다면 개체 하나의 스트링의 길이는 $8\times c\times(p+1)$개다. 최상위 비트부터 $8\times c\times p$개의 스트링은 식(10)과 같이 초구체의 중심으로 디코딩을 행한다.

$$C_{ik} = x_{k\min} + \frac{d}{256}( x_{k\max} - x_{k\min}) \quad (10)$$

$d$는 각 8비트의 스트링을 0에서 255사이의 값으로 디코딩한 값이다.

초구체의 중심은 식(11)에 의해 디코딩한다.

$$R_i = R_{\min} + \frac{d}{256}( R_{\max} - R_{\min}) \quad (11)$$

두 번째로 고려해야 될 사항은 적절한 적합도 평가함수의 선택이다. 본 논문에서 탐색하고자하는 파라미터가 각 군집을 포함하는 최소의 초구체이므로 식(12)와 같이 결정하였다.

$$f = \frac{1}{1 + w_e * E + w_v * V + w_o * O + w_i * I} \quad (12)$$

여기서 $w_e$, $w_v$, $w_o$, $w_i$는 각 파라미터의 가중치가 된다. 각 파라미터의 값의 범위가 다르므로 가중치의 선택은 중요하다.

파라미터 $E$는 어느 초구체에도 포함되지 않는 에러 데이터의 수이다. 이는 가장 중요하게 고려해야 할 요소이다. 에러 데이터의 수는 식(13)으로 판단한다.

$$\sum_{k=1}^{p} ( x_{jk} - C_{ik})^2 \le R_i^2 \quad (13)$$

파라미터 $V$는 초구체의 부피를 나타낸다. 전체 데이터를 둘러싸면서 최소의 부피로 둘러싸기 위해 이 변수의 고려가 필요하다. 전체 부피는 초구체의 반지름에 비례하므로 식(14)에 의해 계산할 수 있다.

$$V = \sum_{i=1}^{c} \prod_{j=1}^{p} R_i \quad (14)$$

이렇게 구한 부피 $V$는 파라미터 $E$에 비해 상당히 크게 되므로 데이터의 개수로 정규화한다.

$$V' = n * \frac{V}{V_{\max} - V_{\min}} \quad (15)$$

파라미터 $O$는 각 초구체가 겹쳤을 때 겹친 부분에 데이터가 존재하는지를 고려하는 변수이고, $I$는 각 초구체가 서로 포함되는지를 고려하는 변수이다. 이 두 변수의 고려로 최소 부피를 가진 초구체로 각 군집의 데이터를 둘러싸는 초구체의 중심을 탐색할 수 있다. 또한, 가중치의 설정도 적절하게 해야 한다.

## 4. 실험 결과

### 4.1 실험 조건

FCM 알고리즘에 있어서 퍼지 분할 행렬을 랜덤수를 발생시켜 초기화했으며, 지수 가중치 $m=2.0$으로 하였으며, 수렴 조건의 역치 $\varepsilon=0.00001$로 하였다.

유전 알고리즘의 경우 전체 개체 집단의 크기는 100으로 하고, 각 유전 인자의 길이는 8비트로 하였다. 단순 유전 알고리즘의 경우 $P_c=1.0$, $P_m=0.1$로 하였다. 적응적 유전 알고리즘의 경우 $k_1=k_3=1.0$, $k_2=k_4=0.5$로 하였다.

### 4.2 실험 결과

실험은 4개의 군집으로 이루어진 임의의 2차원 패턴 데이터로 각각의 군집이 16개, 4개, 9개, 4개의 데이터로 구성되어 있다. 데이터의 분포는 그림 2와 같다.

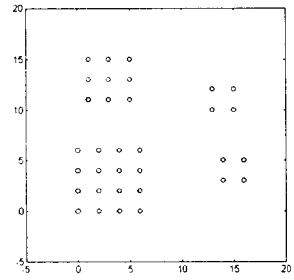
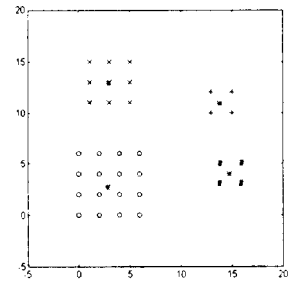
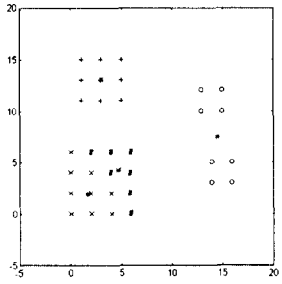


그림 2 데이터 분포

그림 3은 FCM 알고리즘으로 얻어진 하드 군집화를 보여준다. 퍼지 분할 행렬에서 가장 큰 값을 각 군집에 속하는 데이터로 할당하여 각기 다른 모양으로 표현하였고, 각 군집 중심은 "*"로 나타내었다. 여기에서 그림 3(a)는 식(6)의 목적 함수 값이 177.9이고, 그림 3(b)는 이보다 큰 191.3이므로 그림 3(a)가 더 최적의 군집화 결과라고 할 수 있다.



(a)



(b)

그림 3 하드 군집화 결과

본 논문에서 제안한 단순 유전 알고리즘을 통해 각 군집을 포함한 원의 중심과 반지름을 구한 결과는 그림 4와 같다. 각 군집을 포함하는 최적의 결과를 도시한 것이다. "+"는 원의 중심을 나타내며, "*"는 FCM 알고리즘을 통해서 구한 군집 중심이다. 탐색한 결과는 최소의 부피가 되도록 각 군집을 둘러싸는 최소 반지름으로 하는 원으로 표현됨을 볼 수 있다.

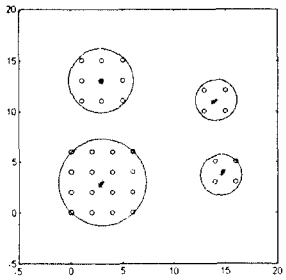


그림 4 단순 유전 알고리즘의 결과

적응적 유전 알고리즘을 통해 구한 결과는 그림 5와 같다. 도시된 결과는 데이터의 분포가 규칙적이므로 FCM 알고리즘으로 구한 군집의 중심을 거의 근사하게 구함을 알 수 있다.

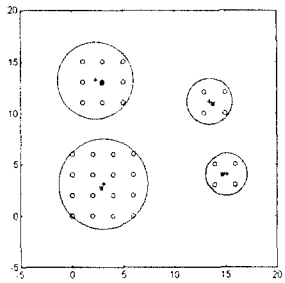


그림 5 적응적 유전 알고리즘의 결과

각각 100회의 실행 결과 기존의 FCM 알고리즘에서는 그림 3(a)처럼 31회의 군집화 결과를 얻었지만, 단순 유전 알고리즘으로 군집을 둘러싸는 원의 중심으로 초기 군집의 중심을 초기화하여 FCM 알고리즘을 수행시키면 91회의 최적 군집화 결과를 얻었고, 적응적 유전 알고리즘의 경우에는 1번만 빼고 모두 군집화에 성공하였다. 즉, 제안한 방법이 기존의 방법보다는 우수하나, 유전 알고리즘이 통계적인 탐색 알고리즘특성상 국부 최소 값에 빠지는 경우도 있었다. 또, 적응적 유전 알고리즘이 간단한 유전 알고리즘보다는 우수한 결과를 얻을 수 있었다.

공간상 표현의 편리성으로 2차원 데이터에 대한 결과만 보였지만 다차원의 패턴 데이터의 경우에도 제안된 방법이 잘 적용될 수 있다.

## 5. 결론

본 논문에서는 유전 알고리즘을 이용하여 FCM 알고리즘의 초기 군집 중심의 선택 방법에 관해 제안하였다. 임의의 패턴 데이터의 군집 분포는 최소 부피의 초구체로 둘러싸일 수 있으며, 이러한 초구체의 여러 파라미터를 유전 알고리즘을 통해 탐색할 수가 있었다. 이렇게 탐색된 초구체의 중심을 기존의 FCM 알고리즘의 초기 군집 중심으로 초기화하는 제안된 방법이 랜덤하게 퍼지 분할 행렬의 계수를 초기화하는 기존의 FCM 알고리즘 방법보다 더 높은 확률로 최적의 군집화를 이룸을 실험을 통해 확인해 보았다. 또한, 단순한 유전 알고리즘보다는 이를 개선한 적응적 유전 알고리즘이 더 높은 확률로 최적해를 찾음을 알 수 있었다.

앞으로의 연구는 각 군집을 최소의 부피로 최적으로 둘러싸는 더 적절한 적합도 평가 함수의 선택 방법, 그리고 국부 최소 값에 더 적은 확률로 수렴하도록 유전 알고리즘의 개선 등이다.

### 참고 문헌


[1] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York, Plenum Press, 1981.

[2] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Reading, MA: Addison-Wesley, 1989.

[3] L. Yao, "Nonparametric learning of decision regions via the genetic algorithm", IEEE Trans. Sys., Man, &Cybern., vol. 26, pp313-321, Apr. 1996.

[4] M. Srinivas, and L. M. Patnaik, "Adaptive probabilities of crossover and mutation in genetic algorithms", IEEE Trans. Trans. Sys., Man, &Cybern., Vol. 24, pp656-667, April, 1994.

[5] J. Liu, and W. Xie, "A Genetics-Based Approach to Fuzzy Clustering", IEEE Trans. Trans. Sys., Man, &Cybern., Vol. 24, pp2233-2240, July, 1995.