진화 프로그램을 이용한 효율적인 퍼지 클러스터링 알고리즘

Effective Fuzzy Clustering Algorithm Using Evolution Program

정 창 호[1], 박 주 영[2], 박 대 희[3]

Changho Jung[1], Jooyoung Park[2], and Daihee Park[3]

[1]고려대 대학원 전산학과, [2]고려대 제어계측 공학과, [3]고려대 전산학과

ABSTRACT

본 논문에서는 기존 FCM(Fuzzy C-Means) 타입 클러스터링 알고리즘의 성능 향상을 위한 설계 방법을 제시한다. 우선 클러스터의 응집성(compactness)과 분리성(separation)을 동시에 고려한 성능 지수를 정의하고, 이를 진화 프로그램을 통하여 최적화 한다. 또한 실험을 통하여 기존 연구들과의 비교 및 제안된 방법론의 유효성을 보인다.

Ⅰ. 서 론

클러스터링은 주어진 데이터 집합의 패턴을 비슷한 성질을 가지는 그룹으로 나누기 위한 방법론으로 이를 위한 많은 알고리즘들이 개발되어 왔다: Rhee등[2]은 새로운 타당성 측정 함수를 제안하고 이를 기반으로한 점증적인 방법의 알고리즘을 제시하였다. Babu등[1]은 최적화 탐색도구인 진화 프로그램이 클러스터링 알고리즘에, 비록 초보적인 단계이지만, 효과적으로 적용될 수 있음을 보였다. 또한 Dave등[4]은 잡음에 강인한 클러스터링 알고리즘의 설계방안을 제시하였다.

본 논문에서는 클러스터의 응집성(compactness)과 분리성(separation)을 동시에 고려한 성능 지수를 진화 프로그램을 통하여 최적함으로써, FCM(Fuzzy C-Means) 타입 클러스터링 알고리즘의 문제점 해결과, 성능 향상을 위한 설계 방법을 제시한다. 특히 제안된 방법론은 첫째, 가능한 중심들로 염색체를 표현함으로써 탐색 공간을 축소한다; 둘째, 밀집 정도(density measure)가 높은 데이터를 클러스터의 예비중심후보로 선택하는 전처리 과정을 통하여 알고리즘의 수렴 속도를 향상한다; 셋째, 진화 프로그램을 통하여 초기화 문제, 클러스터의 개수 문제, 그리고 지역적 최적치(local optimum)등의 문제를 해결한다; 넷째, 잡음 데이터가 중심 집합에 포함될 확률을 줄이고 탐색에 미치는 영향을 최소화함으로써 잡음 데이터에 대하여 강인한 능력을 갖는다.

본 논문의 순서는 다음과 같다: 2장에서는 본 논문을 통해 제안되는 방법론에 대해 설명하고, 3장에서는 실험을 통해 기존 연구와 비교 분석한다. 마지막으로 4장에서는 결론을 제시한다.

Ⅱ. 진화 프로그램을 이용한 퍼지 클러스터링

(1) 진화프로그램의 적용과 성능지수

클러스터의 응집성(compactness)과 분리성(separation)을 동시에 고려한 성능 지수를 진화 프로그램을 통하여 최적화 하는 본 논문의 방법론을 실현하기 위한 기본 단계로써, 본 방법론에 기본적인 개념을 제공하는 BOFCM(Bi-Objective FCM)[3]의 목적 함수를 다음과 같이 새롭게 공식화 한다:

Minimize $\Phi(V)$

and

$$\text{Maximize } L = \sum_{t=1}^{c} \sum_{s<t} \|v_t - v_s\|^2 \tag{1}$$

$$\text{where } \Phi(V) = \min_{U} J(U, V) = \min \sum_{t=1}^{c} \sum_{i=1}^{n} u_{i,t}^{m} \|x_i - v_t\|^2$$

where U : partition matrix,

V : center matrix.

다음 단계로, 식(1)의 v_t와 $u_{i,t}$를 각각 독립 변수와 종속 변수로 설정하고, 단방향성(unidirectional)의 입장에서 식(1)을 최적화 시키는 방법으로써, 중심(v_t)의 탐색은 진화 프로그램으로, 종속 변수인 소속값($u_{i,t}$)의 갱신은 편미분을 이용한다. 이때, 각 개체를 평가하기 위한 성능 지수는 다음과 같이 정의된다:

$$E = \alpha * \frac{1}{\Phi(V)} + \beta * L \tag{2}$$

여기서 α, β는 가중치(weight)이고, $\Phi(V)$와 L은 식(1)에서 정의된 함수로써 응집성과 분리성을 측정하는 평가함수이다.

(2) 수렴속도의 향상과 잡음 처리능력

본 논문에서는 가능한 중심들로 염색체를 표현함으로써 진화 프로그램이 탐색해야할 해공간을 축소한다. 또한, 밀집 정도(density measure)가 높은 데이터를 클러스터의 예비중심후보로 선택하는 전처리 과정을 통하여 가능한 중심들을 선택한다. 이와 같은 방법으로 구해진 예비중심후보를 이용하여 진화 프로그램의 초기 개체 집단 구성과 유전 연산에 사용함으로써 알고리즘의 수렴 속도를 향상시킨다. 입력 데이터의 밀집 정도는 다음과 같다:

$$D_i = \sum_{j=1}^{n} \exp\left(- \frac{\|x_i - x_j\|^2}{(r_a/2)^2} \right) \tag{3}$$

여기서 r_a 는 양수값을 갖는 상수이다.

기존의 FCM 타입 알고리즘의 경우, 소속값 갱신을 위한 식이 갖는 제약조건으로 인하여 잡음 데이터가 매우 높은 소속값을 가질 수 있고, $u_{i,l}$와 중심(v_l)을 상호 종속 변수로 설정함으로써 잡음 데이터가 갖는 소속값이 중심값 측정에 민감하게 작용하는 단점을 가지고 있다[4]. 그러나 본 논문에서는 중심(v_l)을 독립변수로 설정함으로 잡음 데이터가 갖는 소속값은 중심 측정에 영향을 미치지 못한다. 또한 전처리 과정을 통하여 밀집 정도가 상대적으로 낮은 잡음은 예비중심후보로 선택될 확률이 줄어들기 때문에 잡음이 중심값 탐색에 미치는 영향이 최소화된다. 따라서 본 논문의 방법론은 잡음에 대하여 강인한 능력을 갖는다.

Ⅲ. 실험 방법 및 결과

본 논문에서 제안된 방법론의 타당성을 보이기 위하여 2가지 테스트 데이터 집합을 사용하였고, m(weight exponent)은 가장 일반적인 값인 2로 설정하였다. 비교 대상과의 성능 비교를 위하여 제안된 알고리즘을 10회 반복 수행하여 얻어진 결과 값들의 평균을 사용하였다. 실험에서 사용된 진화 프로그램의 매개변수는 Table 1에 나타나있다.

Table 1. 진화 프로그램에 사용된 매개변수

매개변수	값
최대 세대 반복 횟수	20
개체집단의 크기	30
교배 연산자의 확률	0.5
돌연변이 연산자의 확률	0.05

(1) Artificial Data Set A

그림 1과 그림 2는 제안된 방법과 BOFCM 알고리즘이 수행된 후의 중심값을 보여준다. Table 2는 제안된 방법이 기존의 FCM 알고리즘이나 BOFCM 알고리즘보다 성능이 우수함을 보인다.

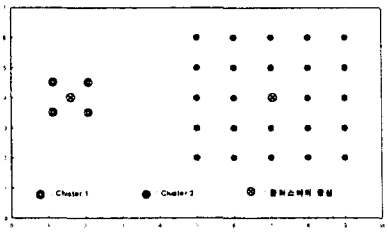

그림 1. 제안된 방법의 결과

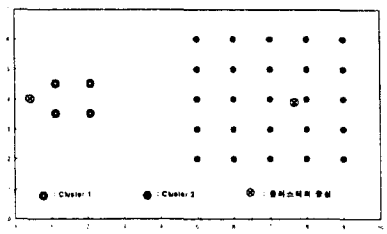

그림 2. BOFCM알고리즘의 결과

Table 2. FCM, BOFCM과의 성능 비교

	FCM	BOFCM	제안된 방법
Misclassification	5	0	0
Ambiguous	0	0	0
PE(Partition Entropy)	0.40	0.31	0.27
S(Xie-Beni)	0.161817	-	0.085568

(2) Artificial Data Set B

그림 3은 잡음 데이터가 포함되지 않은 경우를 보여주고, 그림 4는 잡음 데이터가 포함된 경우의 입력 데이터를 보여주고 있다. Table 3은 제안된 방법론이 잡음에 강인한 면을 갖고 있다는 것을 보여준다.

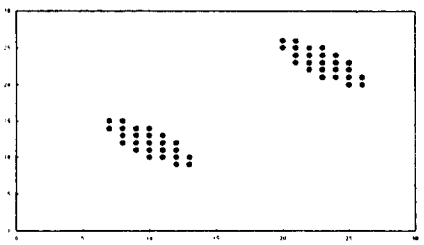

그림 3. 데이터 집합 B

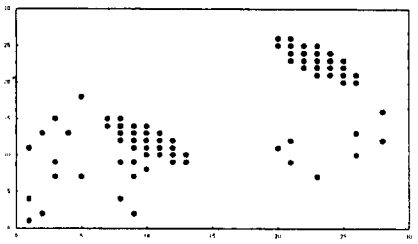

그림 4. 잡음이 포함된 데이터 집합 B

Table 3. 제안된 방법과 FCM 알고리즘과의 성능비교

	잡음 추가전	잡음 추가후
제안된 방법	(10, 12) (23, 23)	(10, 12) (23, 23)
FCM	(10, 12) (22.99, 22.99)	(8.4, 10.5) (23.22, 21.4)

Ⅵ. 결론

본 논문에서는 클러스터의 응집성(compactness)과 분리성(separation)을 동시에 고려한 목적 함수를 진화 프로그램을 통하여 최적화으로써, 기존 FCM(Fuzzy C-Means) 타입 클러스터링 알고리즘의 문제점을 해결하고 성능향상을 위한 설계 방법을 제시 하였다.

Ⅴ. 참고 문헌

1. G. P. Babu and M. N. Murty, "Clustering with evolution strategies," *Pattern Recognition*, vol. 27, no. 2, pp. 321-329, 1994.
2. H. Rhee and K. Oh, "A design and analysis of objective function-based unsupervised neural networks for fuzzy clustering," *Neural Processing Letters*, vol 4, pp. 83-95, 1996.
3. H. Wang, C. Wang and G. Wu, "Bicriteria fuzzy c-means analysis," *Fuzzy Sets and Systems*, vol. 64, pp. 311-319, 1994.
4. R. N. Dave and R. Krishnapuram, "Robust clustering methods: A unified view," *IEEE Trans. on Fuzzy Syst.*, vol. 5, no. 2, pp. 270-293, 1997.
5. Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, 2nd edition, Springer-Verlag, New York, 1994.