노이즈에 강한 밀도를 이용한 Fuzzy C-means 클러스터링 알고리즘

Noise resistant density based Fuzzy C-means Clustering Algorithm

고정원, 최병인, 이정훈
한양대학교 전자 컴퓨터 제어 계측 공학과
E-mail: { jwko, bichoi, frhee }@fuzzy.hanyang.ac.kr

요 약

Fuzzy C-Means(FCM) 알고리즘은 probabilitic 멤버쉽을 사용하는 클러스터링 방법으로서 널리 쓰이고 있다. 하지만 이 방법은 노이즈에 대하여 민감한 성질을 가진다는 단점이 있다. 따라서 본 논문에서는 이러한 노이즈에 민감한 성질을 보완하기 위해서 데이터의 밀도추정을 이용하여 새로운 FCM 알고리즘을 제안한다. 본 논문에서 제안된 알고리즘은 FCM과 비슷한 성능의 클러스터링 수행이 가능하며, 노이즈가 포함된 데이터에서는 FCM보다 더 나은 성능을 보여준다.

Key Words : Fuzzy c-means, Noise, Density estimation, Parzen-window

1. 서 론

Fuzzy C-Means(FCM)은 유클리디안 공간상에서 거리를 이용하여 퍼지 멤버쉽을 할당함으로서 클러스터링을 수행한다[1]. 하지만 노이즈가 섞여 있는 데이터를 클러스터링 하는 경우는 결과가 좋지 않을 수 있다. 이러한 이유는 패턴과 클러스터 센터간 거리에 따라 클러스터들간의 상대적인 멤버쉽을 할당하기 때문이다. 이러한 단점을 보완하기 위하여 PCM이나 CFCM등의 노이즈에 강한 클러스터링 방법들이 제안되었다[2][3]. 하지만 PCM과 같은 경우 결과값이 초기센터에 많은 영향을 받고 또한 클러스터 센터들이 한곳으로 겹치는 현상이 일어나기도 한다[4]. CFCM 역시 초기 센터값에 따라 결과가 안 좋게 나타날수가 있다. 따라서 본 논문에서는 패턴의 밀도추정을 이용한 FCM 클러스터링 방법을 제안하고 있다. 제안하는 방법은 패턴의 밀도 추정값이 작으면 좋지 않은 패턴으로 판단하여 상대적으로 작은 퍼지 멤버쉽을 할당하고, 밀도 추정값이 높으면 좋은 패턴이라 판단하여 높은 멤버쉽을 할당한다. 따라서 노이즈가 포함된 데이터에 대해서는 FCM보다 더 뛰어난 클러스터링이 가능하다. 그리고 각 패턴의 밀도추정에는 Parzen-window 방법을 사용하였다. 본 논문은 다음과 같이 구성된다. 두 번째 절에서는 다른 클러스터링 알고리즘들에 대하여 소개하고 세 번째 절에서는 여기서 제안하는 DBFCM에 대하여 설명한다. 네 번째 절에서는 제안하는 알고리즘과 다른 알고리즘의 실험결과를 비교하고, 마지막으로 결론을 맺겠다.

2. FCM, CFCM 알고리즘

2-1. Fuzzy C-Means (FCM)

일반적으로 잘 알려진 FCM은 다음과 같은 목적함수를 최소화 하는 것을 기초로 한다.

$$J_{FCM} = \sum_{k=1}^{k=n}\sum_{i=1}^{i=c}(u_{ik})^m \| \mathbf{x}_k - \mathbf{v}_i \|^2 \qquad (1)$$

$$\text{subject to } \sum_{i=1}^{i=c} u_{ik} = 1 \qquad (2)$$

여기서 $\mathbf{x}_k$는 k번째 패턴을 나타내고 $\mathbf{v}_i$는 i번째 클러스터 센터를 나타낸다. 그리고 u_{ik}는 패턴k가 클러스터 i에 소속되는 멤버쉽 값을 나타내고, m은 퍼지화 상수로서 $m \in (1, \infty)$인 조건을 만족한다. 위의 목적함수가 최소값을 갖도록 해주는 소속도와 클러스터 센터는 다음의 식에 의해서 구할 수 있다.

$$u_{ik} = \left(\sum_{j=1}^{j=c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}} \right)^{-1} \qquad (3)$$

$$\text{and} \qquad \mathbf{v}_i = \frac{\sum_{k=1}^{k=n} \left(u_{ik}^m \mathbf{x}_k \right)}{\sum_{k=1}^{k=n} \left(u_{ik}^m \right)} \qquad (4)$$

2-2. Credibility Fuzzy C-Means (CFCM)

CFCM은 노이즈에 민감한 FCM의 약점을 보완할수 있는 방법이다. 여기서는 credibility 라는 새로운 변수를 사용하여, 노이즈에 민감한 기존의 FCM의문제점을 극복하고 있다[3].

패턴 $\mathbf{x}_k$의 credibility Ψ_k는 다음과 같다.

$$\Psi_k = 1 - \frac{(1-\theta)\alpha_k}{\max_{j=1\ldots n}(\alpha_j)} \quad , \quad 0 \le \theta \le 1 \qquad (5)$$

$$\text{where,} \qquad \alpha_k = \min_{i=1\ldots c}(d_{ik}) \qquad (6)$$

위의 식과 같이 각 패턴은 데이터군에 가까우면 높은 credibility를 가지고, 멀리 떨어져 있으면 낮은 credibility를 가진다.

CFCM의 목적함수는 다음과 같다.

$$J_{CFCM} = \sum_{k=1}^{k=n}\sum_{i=1}^{i=c}(u_{ik})^m \| \mathbf{x}_k - \mathbf{v}_i \|^2, \qquad (7)$$

$$\text{subject to } \sum_{i=1}^{i=c} u_{ik} = \psi_k \qquad (8)$$

목적함수를 최소화 시켜주는 패턴과 클러스터 간의 멤버쉽은 다음 식에 의해 구할 수 있다.

$$u_{ik} = \frac{\Psi_k}{\sum_{j=1}^{j=c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \qquad (9)$$

$$, \ 1 \le i \le c, \ 1 \le k \le n$$

그리고 클러스터 센터에 대한 업데이트 식은 FCM과 거의 같기 때문에, FCM과 같이 식(4)를 사용하여 클러스터 센터를 구한다.

3. Density Based FCM (DBFCM)

CFCM에서는 패턴과 데이터군의 거리를 계산하여 credibility 라는 변수의 식에 사용함으로서, 멤버쉽 계산에 사용했다. 하지만 본 논문에서 제안하는 방법에서는 패턴과 데이터군 간의 거리 대신 각 패턴의 밀도를 이용한다. 보통 정상적인 패턴은 그 주위에 다른 패턴들이 많이 분포해있지만, 노이즈 패턴의 경우는 그 주위에 다른 패턴들이 없고 홀로 떨어져 있을 것이다. 따라서 밀도가 높은 패턴에 대해서는 높은 멤버쉽을 할당하고, 밀도가 낮은 패턴에 대해서는 상대적으로 낮은 멤버쉽을 할당한다. 그리고 패턴의 밀도추정 방법으로는 Parzen-window를 사용한다.

3-1. Parzen-window method

패턴$\mathbf{x}_k$을 포함하는 영역R이 있을 때, 고정된 크기의 영역R안에 존재하는 패턴들의 개수를 계산하여 밀도 추정값 $p_k(\mathbf{x})$를 얻을수있다. 이때 다음의 식을 이용하여 $p_k(\mathbf{x})$를 구한다[5].

$$p_k(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \qquad (10)$$

h : 다차원 영역 R의 선분길이

$V = h^d$: 다차원 영역 R의 부피

n : 데이터의 개수

이때 window 함수 $\varphi(u)$로는 일반적으로 많이 쓰는 가우시안 모양을 사용한다.

$$\varphi(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2} \qquad (11)$$

$p_k(\mathbf{x})$값은 smoothing parameter인 h에 영향을 받는다. h가 커지면 데이터의 밀도분포가 완만해지고,h가 작아지면 밀도분포가 날카로워진다. 따라서 적절한 h값을 선택하는것이 중요하다.h값은 데이터의 표준편차와 연관이 있기 때문에, 다음과 같이 적절한 h값을 구한다[6].

$$h = \sigma\left(\frac{4}{(m+2)n}\right)^{\frac{1}{m+4}} \qquad (12)$$

n : 데이터의 개수

m: 벡터공간의 차수

σ : 데이터의 표준편차

3-2. Proposed DBFCM

식(10),(11),(12)를 사용하여 $p_k(\mathbf{x})$를 구한다. 이때 DBFCM은 다음의 목적함수를 가진다.

$$J_{DEFCM} = \sum_{i=1}^{i=c}\sum_{k=1}^{k=n} u_{ik}^m \| \mathbf{x}_k - \mathbf{v}_i \|^2, \qquad (13)$$

$$\text{subject to} \quad \sum_{i=1}^{i=c} u_{ik} = p_k \qquad (14)$$

목적함수를 최소화 하는 멤버쉽은 다음의 식으로 구할 수 있다.

$$u_{ik} = \frac{p_k}{\sum_{j=1}^{j=c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \qquad (15)$$

$$, \ 1 \le i \le c, \ 1 \le k \le n$$

CFCM과 마찬가지로 클러스터 센터에 대한 업데이트 식은 FCM과 거의 같기 때문에, 역시 식(4)를 사용하여 클러스터 센터를 구한다.

Density Based Fuzzy C-Means Clustering

Step1. Initilization

Initialize random k center $\mathbf{v}_k$ and m

Compute the initial membership u_{ik}using (3)

Step2. Minimization of the objective function

Do :

Update $\mathbf{v}^t$ using (4)

Compute u_{ik}^t using (18)

Until : $u_{ij}^{(t+1)} - u_{ij}^{(t)} < \epsilon$

4. 시뮬레이션 결과

본 절에서는 본 논문의 타당성을 보이기 위해 FCM, CFCM, 그리고 제안하는 DBFCM 알고리즘으로 데이터들을 실험 후, 그 결과를 비교해 보겠다. 모든 실험에 fuzziness parameter인 m값은 2를 사용하였다.

4-1. Spherical Data

이번 실험에서 구형의 두 개의 클러스터로 나눠지는 데이터를 클러스터링 하였다. 클러스터당 27개씩 총 54개의 패턴으로 이루어진 데이터이다. 그 결과 FCM, CFCM, DBFCM 방법 모두 좋은 결과를 나타내고 있다.

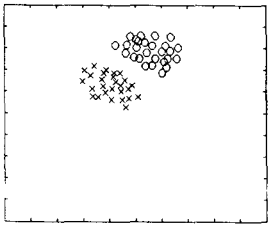
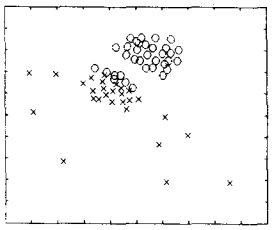


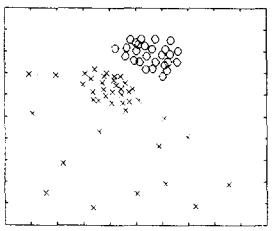
그림 1. 노이즈가 없는 구형의 데이터의 결과 : FCM, CFCM, DBFCM

4-2. Spherical Data with Noises

그럼 이번에는 앞의 데이터에 14개의 노이즈를 포함시켰을 경우 결과를 보자. 노이즈에 민감한 FCM은 역시 아래쪽에 위치한 노이즈들에 영향을 받아 두 클러스터가 만나는 지점에서 결과가 좋지 않게 나왔다. 하지만 CFCM과 DBFCM의 경우에는 노이즈에 영향을 받지 않고 좋은 클러스터링 결과를 보여준다.



(a)

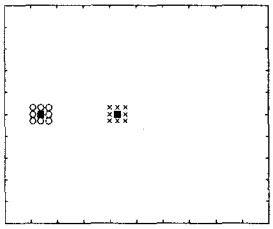


(b)

그림 2. 노이즈가 포함된 구형데이터의 클러스터링 결과 (a) FCM (b) CFCM, DBFCM

4-3. Same volume "squares" Data

같은 부피를 가진 두 개의 클러스터로 나뉘어지는 데이터를 FCM, CFCM, DBFCM 방법으로 클러스터링을 하여 결과를 비교해본다. 실험 후 그림과 같은 결과를 볼 수 있는데 세가지 방법 모두 올바른 클러스터 센터를 얻을수 있었다.



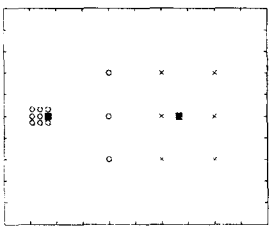
■ : 최종 클러스터 센터

그림 3. 같은 부피의 두 개의 사각형 모양의 데이터 클러스터링 결과 : FCM, CFCM, DBFCM

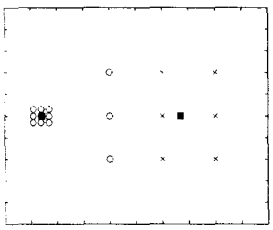
4-4. Different volume "squares" Data

이번엔 같은 개수, 다른 부피를 가진 사각형 모양의 데이터를 같은 방법으로 클러스터링 해보았다.

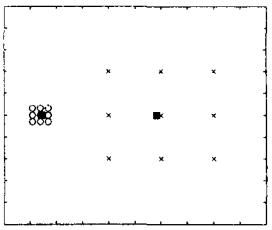
그 결과 FCM에서는 클러스터 센터가 오른쪽으로 치우치는 현상이 나타난다. 때문에 큰 사각형의 왼쪽에 위치한 패턴 세개는 작은 사각형 쪽으로 잘못 클러스터링 된 것을 볼 수 있다. CFCM의 경우는 FCM보다는 결과가 나았지만 역시 올바른 클러스터 센터를 찾지 못하고 역시 세 패턴이 잘못 분류 되었다. 하지만 DBFCM의 경우에는 가장 올바른 클러스터 센터를 찾을 수 있었고 때문에 모든 패턴들을 올바르게 분류할 수 있었다.



(a)



(b)



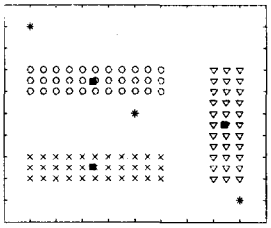
(c)

■ : 최종 클러스터 센터

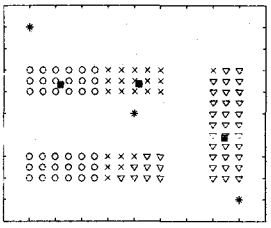
그림 4. 다른 부피의 두 개의 사각형 모양의 데이터 결과 (a) FCM (b) CFCM (c) DBFCM

4-5. 3-Lines Data

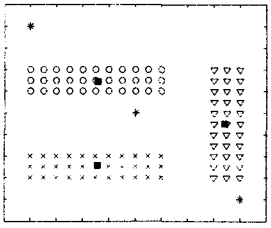
이번엔 라인당 33개씩 총 99개의 패턴을 가진 라인 데이터의 결과를 보겠다. 실험 결과 FCM과 DBFCM은 초기 클러스터의 위치에 관계없이 올바른 클러스터링을 수행 하지만 CFCM의 경우 초기 클러스터 센터를 잘못 주었을 경우 좋지 않은 결과를 보여준다. 이는 CFCM에서 불안정한 초기센터가 credibility에 영향을 주면서 수행이 되기 때문이다.



(a)



(b)



(c)

* : 초기 클러스터 센터

■ : 최종 클러스터 센터

그림 5. 세 개의 선 모양의 데이터 결과 (a) FCM (b) CFCM (c) DBFCM

5. 결 론

본 논문에서는 Parzen-window방법을 사용한 DBFCM 알고리즘을 제안하였다. 결과에서 보았듯이 제안하는 알고리즘은 기본적으로 FCM과 같은 성능을 보이면서, 노이즈가 포함된 데이터에 대해서는 FCM 보다 우수한 성능을 보였다. 그리고 CFCM과 달리 초기 센터값에 민감하다는 단점을 보이지도 않는다. 하지만 제안하는 DBFCM 알고리즘에서는 원하는 결과값을 얻기 위해, 최적의 smoothing parameter h값을 구하는 것이 매우 중요하다. 따라서 어떠한 형태의 데이터가 입력되더라도, 최적의 h값을 구할 수 있는 방법을 찾는 것이, 향후 해결해야 할 과제이다.

감사의 글 : 본 연구는 한국과학기술원 영상정보특화연구센터를 통한 국방과학연구소의 연구비 지원으로 수행되었으며 연구비 지원에 감사 드립니다.

참 고 문 헌

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

[2] R. Krishnapuram and J. Keller, "The Possibilistic Approach to Clustering," *IEEE Transactions on Fuzzy Systems*, Vol. 1, No. 2, pp. 98-110, 1993.

[3] K.Chintalapudi and M.Kam,"A noise resistant fuzzy c means algorithm for clustering," IEEE Conference On Fuzzy Systems Proceedings, May 1998.

[4] N.Pal, K.Pal, and J.Bezdek, "A Mixed c Means Clustering Model," *IEEE Int. Conf. Fuzzy Systems*, 1997, Spain, pp. 11-21.

[5] H.Duda, P.Hart, and D. stork, *Pattern Classification*, second edd. John Wiley &Sons, 2001.

[6] A.Bowman and A.Azzalini, *Applied Smoothing Techniques for Data Analysis*, London : Oxford University Press, 1997.