Journal of The Korea Society of Computer and Information
Vol. 20 No. 12, pp. 15-20, December 2015

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.12.015

An Improved Clustering Method with Cluster Density Independence

Byeong-Hyeon Yoo *, Wan-Woo Kim **, Gyeongyong Heo ***

Abstract

In this paper, we propose a modified fuzzy clustering algorithm which can overcome the center deviation due to the Euclidean distance commonly used in fuzzy clustering. Among fuzzy clustering methods, Fuzzy C-Means (FCM) is the most well-known clustering algorithm and has been widely applied to various problems successfully. In FCM, however, cluster centers tend leaning to high density clusters because the Euclidean distance measure forces high density cluster to make more contribution to clustering result. Proposed is an enhanced algorithm which modifies the objective function of FCM by adding a center-scattering term to make centers not to be close due to the cluster density. The proposed method converges more to real centers with small number of iterations compared to FCM. All the strengths can be verified with experimental results.

▸ Keyword : Clustering, FCM, Cluster density, Density independence

I. Introduction

클러스터링은 주어진 데이터를 유사성에 기준하여 몇 개의 그룹으로 나누는 비교사(unsupervised) 학습 방법 중 하나로 패턴인식의 주요 기법 중 하나이다. 소속도 함수(membership function)에 의해 부분 소속도를 나타내는 퍼지 집합이 Zadeh에 의해 소개된 이후, 퍼지 집합은 클러스터링 분야에 도입되었고 퍼지 클러스터링은 대표적인 클러스터링 기법 중 하나로 자리 잡았다. Bezdek[1]에 의해 일반화된 Fuzzy C-Means(FCM)는 퍼지 클러스터링 중 가장 널리 사용되는 방법 중 하나이다. FCM은 간단하면서도 효과적인 클러스터링 방법이지만, 구해진 소속도가 직관적인 값과 일치하지 않는 경우가 있으며, 잡음이 많은 환경이나 클러스터의 중심이 인접하고 클러스터의 밀도가 서로 다른 경우 밀도가 높은 클러스터 쪽으로 클러스터 중심이 쏠리는 현상이 발생한다. 클러스터 밀도가 높은 쪽으로 클러스터의 중심이 쏠리는 현상은 FCM에서 유클리드 거리 또는 그 변형을 사용하기 때문으로, FCM의 목적함수에서는 데이터 개수가 많은 클러스터가 데이터 개수가 적은 클러스터에 비해 목적함수에 더 많은 영향을 미치기 때문이다.

이 논문에서는 클러스터의 밀도 차이로 인해 클러스터 중심이 치우치는 현상을 해결할 수 있는 새로운 클러스터링 방법을 제안한다. 제안하는 '클러스터 밀도에 무관한 클러스터링 알고리듬 (Density-Independent FCM, DI-FCM)'은 기존의 FCM 알고리즘이 갖는 클러스터 밀도 차이에 대한 민감성을 줄이기 위해 클러스터 중심 사이의 거리 합을 나타내는 항을 FCM의 목적함수에 추가하였다. FCM 알고리즘에서 클러스터 간 밀도 차이로 인하여 밀도가 높은 클러스터 쪽으로 중심이 쏠리는 현상은 클러스터의 중심 사이의 거리를 가능한 멀게 함으로써 밀도 차이에 의해 발생하는 클러스터 중심의 치우침 현상을 개선할 수 있다.

이 논문의 구성은 다음과 같다. 먼저 2장에서는 기존의 퍼지 클러스터링 기법에 대하여 설명한다. 3장에서는 제안하는 방법을 소개하고, 4장 실험 및 결과를 통하여 기존의 방법과 비교하여 제안하는 방법의 유효성을 보인다. 마지막 5장에서 4장의 실험결과를 바탕으로 결론 및 향후 연구방향에 대해서 언급한다.

• First Author: Byeong-Hyeon Yoo, Corresponding Author: Gyeongyong Heo
*Byeong-Hyeon You (youyooo@naver.com), Dept. of Electronic Engineering, Dong-eui University
**Wan-Woo Kim (wwkim614@naver.com), Dept. of Electronic Engineering, Dong-eui University
***Gyeongyong Heo (hgycap@deu.ac.kr), Dept. of Electronic Engineering, Dong-eui University
• Received: 2015. 11. 19, Revised: 2015. 11. 30, Accepted: 2015. 12. 18.
• This work was supported by 2015 Dong-eui University Research Grant(2015AA023).

II. Preliminaries

1. Fuzzy C-Means(FCM)

클러스터링은 유사성의 개념을 바탕으로 데이터를 특정 클러스터에 소속되도록 하는 방법이며, 퍼지 클러스터링은 퍼지의 불완전한 소속도를 도입하여 데이터가 하나 이상의 클러스터에 부분적으로 소속될 수 있도록 하는 방법이다. 클러스터링은 역사가 오랜 만큼 다양한 알고리즘이 제안되었고, 대표적인 방법 중 하나가 Bezdek이 제안한 FCM이다. FCM은 제약 조건이 있는 최적화 문제(constrained optimization problem)[2][3]로 표현할 수 있다. n개의 d차원 데이터 $X=\{x_i|1 \le k \le n, x_k \in R^d\}$가 주어졌을 때, 이를 c개 그룹으로 클러스터링하기 위해서는 식 (1)의 목적 함수를 최소로 하여야 한다.

$$J_{FCM}=\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}\| v_i-x_k \|_A^2 = \sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}d_{ik}^2 \quad (1)$$

이 때 μ_{ik}는 k번째 데이터 포인트 x_k가 i번째 클러스터에 소속되는 정도를 나타내는 소속도 값(membership value)을, v_i는 i번째 클러스터의 중심을, $m(1 < m < \infty)$은 퍼지화 정도를 나타내는 상수(fuzzifier constant)로 일반적으로 2로 설정된다. d_{ik}는 k번째 데이터 포인트와 i번째 클러스터 중심 사이의 거리를 나타낸다. 이 논문에서는 유클리드 거리(euclidean distance)를 사용하였다.

FCM에서 하나의 데이터 포인트 x_k는 c개의 클러스터에 소속될 수 있지만, 각 클러스터에 소속되는 정도는 서로 달라지며, c개 클러스터에 소속되는 정도는 전체 소속도의 합이 1이 되어야 한다는 제약조건(constraint)을 만족시켜야 한다.

$$\sum_{i=1}^{c}\mu_{ik}=1 \quad (2)$$

퍼지 클러스터링은 식 (1)의 목적 함수는 반복 최적화 알고리즘을 이용하여 최적화하며, 라그랑즈 승수법(Lagrange multiplier method)을 이용하여 식 (3)과 식 (4)의 갱신 식(update equation)을 얻을 수 있다.

$$v_k=\frac{\sum_{i=1}^{N}\mu_{ik}^{m}x_i}{\sum_{i=1}^{N}\mu_{ik}^{m}} \quad (3)$$

$$\mu_{ik}=\frac{\frac{1}{\| v_i-x_k \|^2}}{\sum_{i=1}^{c}\frac{1}{\| v_i-x_k \|^2}} \quad (4)$$

FCM은 처음 소개된 이후 원형 그대로 또는 주어진 문제에 맞게 변형된 형태로 많은 문제에 성공적으로 사용되어 왔지만 다양한 형태의 변형이 존재한다는 것은 FCM이 모든 문제에 완벽한 것은 아니라는 방증이기도 하다. 이 논문 또한 FCM의 문제점에서부터 시작하여 이를 해결하기 위한 변형된 형태의 퍼지 클러스터링을 제안한다.

III. Density Independent Fuzzy C-Means

FCM에서 밀도가 다른 클러스터가 존재하는 경우 클러스터 중심이 밀도가 높은 클러스터 쪽으로 치우치는 현상은 FCM이 클러스터 중심에서 데이터까지의 유클리드 거리 또는 그 변형에 기반하고 있기 때문이다[4][5]. 이처럼 클러스터 중심이 쏠리는 현상을 방지하기 위해 이 논문에서는 클러스터의 중심이 가능한 멀리 떨어져 있도록 하는 항을 FCM의 목적함수에 추가하였다. 제안하는 변형된 FCM의 목적함수는 식 (5)와 같다.

$$J_{DI}(V, U|X)=\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}\| v_i-x_k \|^2 - \alpha\sum_{a=1}^{c}\sum_{b=1}^{c}\| v_a-v_b \|^2 \quad (5)$$

이 때 $\alpha(>0)$는 중심이 떨어져 있는 정도를 목적함수에 반영하는 비율을 나타내는 상수이다. m은 FCM의 목적함수에서와 같이 퍼지화 정도를 나타내는 값으로 일반적으로 사용되는 2를 사용하였다.

식 (5)가 식 (1)의 FCM의 목적함수와 다른 점은 두 번째 항이다. 두 번째 항에서는 c개 클러스터 중심 사이의 거리 합을 계산한다. α는 양의 값을 가지고, J는 최소값을 가지도록 최적화한다. 따라서 클러스터 중심 사이의 거리 합이 클수록 목적함수 J는 작은 값을 가지게 된다. 식 (5)을 라그랑즈 승수법을 사용하여 갱신식을 구할 수 있다. 먼저 J를 u_{ik}로 편미분하면 식 (6)을 얻을 수 있다.

$$\frac{\partial J_{DI}}{\partial \mu_{ik}}=2\mu_{ik}\| v_i-x_k \|^2-\lambda_i \quad (6)$$

식(6)을 μ_{ik}에 대해 정리하면 식 (7)과 같이 나타낼 수 있다.

$$\mu_{ik} = \frac{\lambda_i}{2 \| v_i - x_k \|^2} \tag{7}$$

DI-FCM 또한 FCM의 제약조건인 식 (2)를 만족하여야 하므로 식(7)을 식(2)에 대입하여 정리하면 식 (8)이 나타낼 수 있다.

$$\frac{\lambda_i}{2} = \frac{1}{\sum_{i=1}^{C} \frac{1}{\| v_i - x_k \|^2}} \tag{8}$$

식(8)을 정리하여 λ_i를 구해 식 (7)에 대입하면 k번째 데이터 포인트가 i번째 클러스터에 소속되는 정도를 나타내는 소속도 갱신식인 식 (9)를 얻을 수 있다.

$$\mu_{ik} = \frac{\frac{1}{\| v_i - x_k \|^2}}{\sum_{i=1}^{c} \frac{1}{\| v_i - x_k \|^2}} \tag{9}$$

DI-FCM을 위한 소속도 갱신식인 식 (9)는 기존 FCM의 소속도 갱신 식인 식 (4)와 동일하다. 즉, DI-FCM에서 소속도는 FCM에서와 마찬가지로 클러스터 중심과 데이터 포인트 사이의 유클리드 거리를 바탕으로 결정된다.

클러스터 중심의 갱신식을 구하기 위해 식 (5)를 v_i에 대해 편미분하면 식 (10)을 얻을 수 있다.

$$\frac{\partial J_{DI}}{\partial v_i} = 2 \sum_{k=1}^{N} \mu_{ik}^2 \| v_i - x_k \| - 2\alpha \sum_{b=1}^{C} \| v_i - v_b \| \tag{10}$$

식(10)을 v_i에 대해 정리하면 클러스터 중심을 업데이트하기 위한 갱신식인 식 (11)을 얻을 수 있다.

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_i - \alpha \sum_{b=1}^{c} v_b}{\sum_{k=1}^{n} u_{ik}^m - \alpha c} \tag{11}$$

클러스터 중심을 나타내는 갱신식인 식 (11)을 FCM의 클러스터 중심 갱신식인 식 (3)과 비교해 보면 분모에는 클러스터의 중심의 개수가, 분자에는 클러스터 중심의 합을 나타내는 항이 각각 추가 되어 있으며 클러스터 중심이 가능한 멀리 위치하도록 α에 의해 조절되고 있음을 알 수 있다.

식 (9)와 식 (11)의 두 갱신 식을 이용하여 제안하는 DI-FCM 알고리듬을 나타내면 Fig. 1과 같이 나타낼 수 있다.

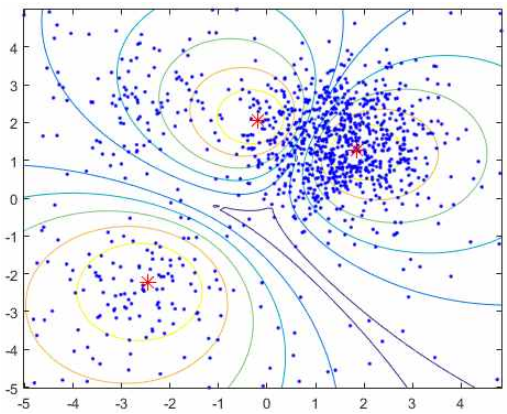
```
1: initialize V
2: initialize t = 0
3: do
4:     t ← t + 1
5:     calculate U using Eq. (9)
6:     calculate V  using Eq. (11)
7: while U and V do not satisfy convergence
   qualification
8: return U and V
```

Fig. 1. DI-FCM Algorithm

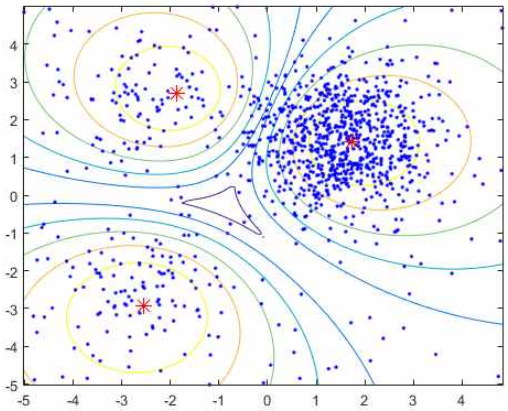
IV. Experimental Results

기존에 다양한 FCM의 변형이 존재하지만 밀도 차이에 의한 클러스터링 오류를 해결하기 위한 클러스터링 방법은 찾기 어려우며, 제안한 방법은 FCM의 변형에도 적용이 가능하므로 실험에서는 제안한 방법을 FCM과 비교하였다.

제안한 방법에서 소속도 μ_{ik}를 구하는 식 (9)는 기존의 FCM과 동일하지만, 클러스터의 중심을 구하는 식 (11)에서 기존 FCM의 차이가 있다. Fig. 2-(a)는 샘플 데이터 집합에 FCM을 적용한 경우의 클러스터링 결과로, 클러스터의 중심이 밀도가 높은 클러스터 쪽으로 치우치고 있음을 알 수 있다. 반면 Fig. 2-(b)의 DI-FCM을 적용한 경우의 클러스터링 결과로 클러스터의 밀도 차이에도 불구하고 정확한 클러스터 중심을 찾아내고 있다. 샘플 데이터는 3개의 클러스터로 구성되어 있으며 오른쪽 위의 밀도가 높은 클러스터는 500개의 데이터 포인트를, 나머지 2개의 클러스터는 100개의 데이터 포인트를 가진다.



(a) FCM



(b) DI-FCM

Fig. 2. Clustering results with dataset I
(Top : FCM, Bottom : DI-FCM)

DI-FCM에서는 중심이 떨어져 있는 정도를 나타내는 상수인 α에 따라 클러스터 중심 사이 거리가 목적함수에 반영되는 정도가 달라진다. 따라서 α 값은 데이터의 분포에 따라 달리 설정되어야 한다. Fig. 3은 Fig. 2의 샘플 데이터 집합에 대해 α를 0에서 8까지 0.2 간격으로 변화시키면서 500번 반복하여 평균 오류를 계산한 것이다. 오류는 데이터 생성에 사용된 실제 클러스터의 중심과 클러스터링 결과로 얻은 클러스터 중심 사이의 유클리드 거리를 합한 것으로 하였다.

$$Error = \sum_{i=1}^{c} \|v_{r,i} - v_{c,i}\|^2 \tag{12}$$

식 (12)에서 $v_{r,i}$는 샘플 데이터 생성에 사용된 클러스터의 중심 좌표를, $v_{c,i}$는 클러스터링 결과로 얻은 클러스터의 중심 좌표를 나타낸다.

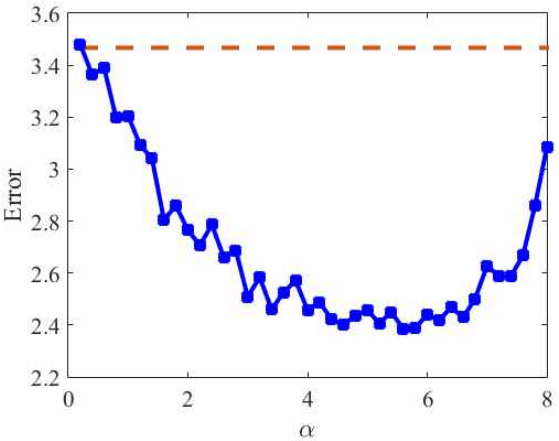


Fig. 3. Clustering error with respect to α on dataset I

Fig. 3에서 알 수 있듯이 DI-FCM은 3.48 정도의 평균 오류 값을 나타내는 FCM에 비해 평균 오류가 적었으며 특히 샘플 데이터의 경우 α가 5.6일 때 최소 평균 오류 2.39를 나타내어 FCM에 비해 오류가 약 31% 줄어든 것을 확인할 수 있다. 하지만 이후 α 값이 커지며 오류 값이 상승하게 되며 이는 클러스터의 중심이 원래 위치보다 너무 멀어지기 때문에 나타나는 현상이다.

Fig. 4는 최적화된 중심을 찾아내기 위한 평균 수렴 속도를 나타낸 것이다. Fig. 4에서 알 수 있듯이 DI-FCM의 수렴 속도는 α가 5.6일 때 72.22로 FCM의 수렴 속도에 비해 약 10% 감소하였다.

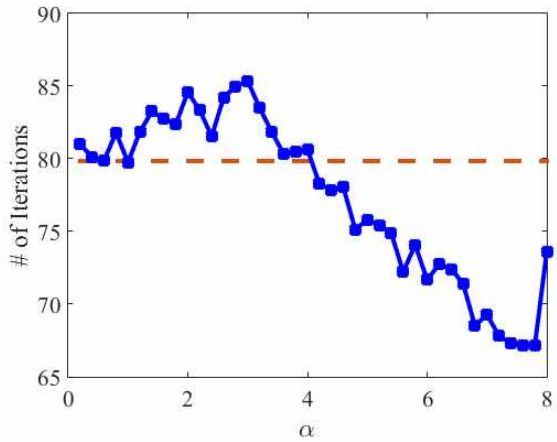
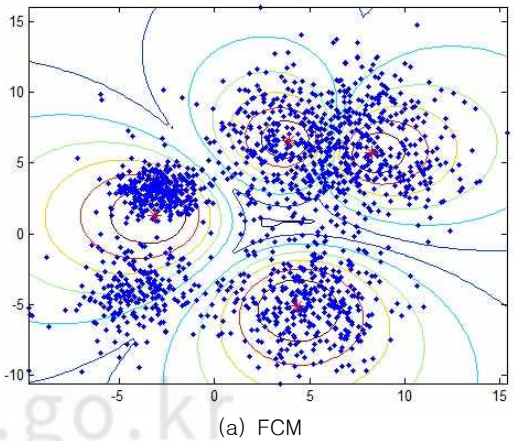


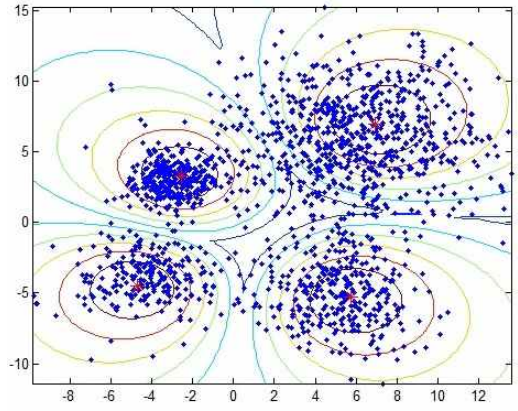
Fig. 4. Convergence speed with respect to α
on dataset I

다음은 서로 다른 밀도를 갖는 4개의 클러스터로 구성된 데이터에 대해 실험하였을 때의 결과이다. 샘플 데이터는 4개의 클러스터로 구성되며 오른쪽 위 클러스터는 700개, 왼쪽 위 클러스터는 300개, 나머지 아래 쪽 두 클러스터는 200개의 데이터 포인트로 구성되었다.

Fig. 5-(a)는 샘플 데이터에 FCM을 적용한 결과이다. Fig. 2-(a)에서와 마찬가지로 밀도가 높은 곳으로 클러스터의 중심이 치우치는 것을 확인할 수 있다. 반면 Fig. 5-(b)는 DI-FCM을 적용한 결과로 클러스터의 수와 상관없이 밀도에 의한 클러스터 중심의 치우침 현상이 줄어드는 것을 확인할 수 있다.



(a) FCM



(b) DI-FCM

Fig. 5. Clustering results with dataset II
(Top : FCM, Bottom : DI-FCM)

Fig. 6은 Fig. 5의 샘플 데이터 집합에 대해 앞에서와 마찬가지로 α를 0에서 8까지 0.2 간격으로 변화시키면서 500번 반복 실험하여 오류와 수렴 속도를 평균하여 나타낸 것이다.

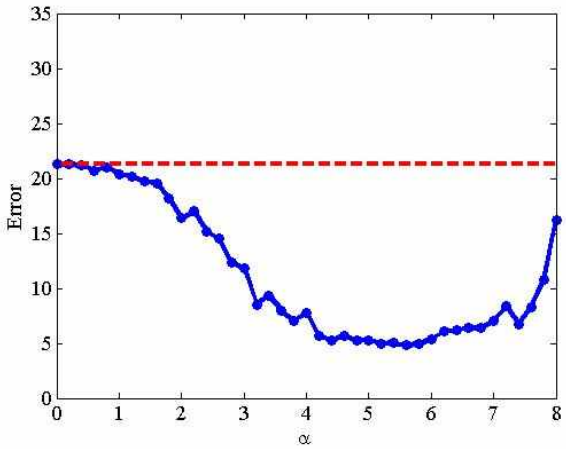


Fig. 6. Clustering error with respect to α on dataset II

Fig. 6을 살펴보면 Fig. 3에서와 마찬가지로 기존의 FCM에 비해 오류 값이 작은 것을 확인할 수 있다. 특히 α값이 5.8일 때 4.83의 오류값을 가져 FCM의 21.26에 비해 약 74% 오류가 줄어들었다.

Fig. 7은 클러스터링의 수렴 속도를 나타낸 값이다. Fig. 7에서는 α값이 5.8일 때 평균 수렴속도 26.38로 FCM의 106.45에 비해 작은 값을 가진다.

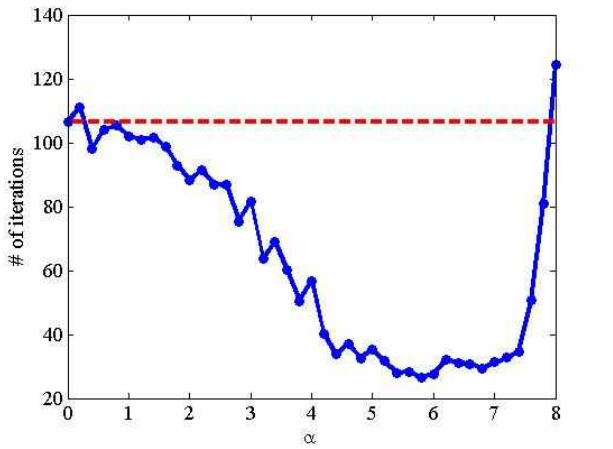


Fig. 7. Convergence speed with respect to α on dataset II

Fig. 6과 7에서 알 수 있듯이 α값이 증가함에 따라 오류는 감소하고 수렴 속도는 빨라진다. 하지만 특정 α값보다 커지는 경우 클러스터 중심이 지나치게 멀리 떨어져 있게 됨으로써 오류는 증가하고 수렴 속도 역시 느려지게 된다.

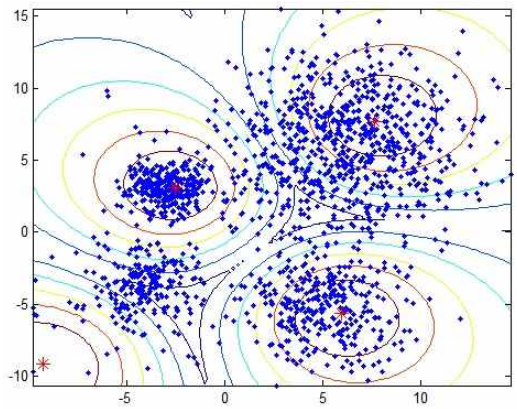


Fig. 8. Deviated center when α=8 with dataset II

Fig. 8은 두 번째 샘플 데이터에서 α값이 8일 때의 클러스터링 결과를 나타낸다. 기존 FCM은 밀도가 높은 쪽으로 클러스터 중심이 쏠리는 현상이 생겼다면, DI-FCM은 α값에 따라 밀도가 낮은 클러스터의 중심이 밀도가 높은 클러스터 중심으로부터 지나치게 멀어지는 현상이 발생함을 알 수 있다.

V. Conclusions

이 논문에서는 클러스터의 밀도 차이로 인해 클러스터 중심이 왜곡되는 현상을 해결하기 위해 클러스터 중심 사이의 거리를 FCM의 목적함수에 추가한 새로운 클러스터링 방법을 제안하였다. 실험 결과에서 제시한 바와 같이 제안하는 DI-FCM (Density Independent FCM)은 기존 FCM에 비해 실제 클러스터 중심에 수렴하는 확률이 높고 수렴 속도 역시 FCM에 비해

빠른 것을 알 수 있다. 하지만 오류 및 수렴 속도가 최적화 되는 α 값은 데이터 집합에 따라 달라지므로 데이터의 특성에 따라 효과적으로 결정하는 방법이 필요하며, 이는 현재 연구 중에 있다.

이 논문에서는 거리 척도로 유클리드 거리를 적용하였다. 하지만 유클리드 거리는 잡음에 민감하며 원형 클러스터만을 다룰 수 있다는 단점이 존재한다. 비록 소속도의 사용으로 일부 완화되지만 국부 최적해에 빠질 수 있다. 따라서 잡음에 대처하기 위해서 regularization[6]의 도입이나 노이즈 클러스터링(noise clustering)[7]의 사용 등을 고려할 필요가 있으며, 클러스터의 다양한 형태에 대응하기 위하여 Gustafson과 Kessel(GK)에 의해 제안된 마할라노비스 거리(Mahalanovis distance)[8][9]를 사용하는 방법이나 Gath와 Geva(GG)[10]에 의해 제안된 방법인 가우스 분포 함수에 반비례하는 값을 거리로 사용하는 것도 고려해 볼 수 있다. 또한 이들 변형에 밀도 무관항을 추가하는 방법을 고려할 수 있으며 이 역시 현재 연구 중에 있다.

REFERENCES

[1] J. Bezdek, Pattern Recognition with fuzzy Objective Function Algorithms, New York, Springer, January 1981.

[2] Sadaaki Muyamoto, Fuzzy Clustering - Basic Ideas and Overview, Handbook of Computational Intelligence, Springer, pp. 293-248, May 2015.

[3] Janmenjoy Nayak, Fuzzy C-means(FCM) Clustering Algorithm: A Decade Review from 2000 to 2014, Systems and Technologies, Vol. 32, No. 2, pp. 133-179, December 2014.

[4] Zarita Zainuddin, An effective Fuzzy C-Means algorithm based on symmetry similarity approach, Applied Soft Computing, Vol. 35, No. 10, pp. 433-448, October 2015.

[5] Basel Abu-Jamous, Fuzzy Clustering, Integrative Cluster Analysis in Bioinformatics, Chapter 13, April 2015.

[6] G. Heo, An Extension of Possibilistic Fuzzy C-Means using Regularization, Journal of the Korean Society of Computer Information, Vol. 15, No. 1, pp. 43-50, January 2010.

[7] R. N. Dave, Characterization and detection of noise in clustering, Pattern Recognition Letters, Vol. 12, No. 11, pp. 657-664, November 1991.

[8] R. Babuska, P. J. van der Veen and U Kaymak, Improved Covariance Estimation for Gustafson-Kessel Clustering, Proceeding of the 2002 IEEE International Conference on Fuzzy Systems, pp. 1081-1085, May 2002.

[9] G. Heo, Extension of the Possibilistic Fuzzy C-Means Clustering Algorithm, Proceedings of KFIS Autumn Conference, Vol. 17, No. 2, November 2007.

[10] I. Gath and A. B. Geva, Unsupervised Optimal Fuzzy Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11, No. 7, pp. 773-791, July 1989.

Authors

Byeong-Hyeon Yoo received the B.S. in Electronic Engineering from Dong-eui University, Korea, in 2015.
He is doing M.S. degrees in Electronic Engineering from Dong-eui University.
His interest includes artificial intelligence and pattern recognition.

Wan-Woo Kim received the B.S. in Electronic Engineering from Dong-eui University, Korea, in 2015.
He is doing M.S. degrees in Electronic Engineering from Dong-eui University.
His interest includes artificial intelligence and pattern recognition.

Gyeongyong Heo received the B.S. and M.S. degrees in Electronic Engineering from Yonsei University and Ph.D. degrees in Computer and Information Science and Engineering from University of Florida, USA, in 2009. Dr. Heo joined the faculty of the Department of Electronic Engineering at Dong-eui University, Busan, Korea, in 2012. His interest includes artificial intelligence, pattern recognition, and robotics.