커널 밀도 추정을 이용한 Fuzzy C-Means의 초기화

허경용* · 김광백**

Initialization of Fuzzy C-Means Using Kernel Density Estimation

Gyeongyong Heo* · Kwang-Baek Kim**

요 약

Fuzzy C-Means (FCM)는 군집화를 위해 널리 사용되는 알고리듬 중 하나로 다양한 응용 분야에서 성공적으로 사용되어 왔다. 하지만 FCM은 여러 가지 단점을 가지고 있으며 초기 원형 설정이 그 중 하나이다. FCM은 국부 최적해에 수렴하므로 초기 원형 설정에 따라 군집화의 결과가 달라진다. 따라서 초기 원형의 설정은 군집화 결과 향상을 위해 중요하다. 이 논문에서는 이러한 FCM의 초기 원형 설정 문제를 해결하는 방안으로 커널 밀도 추정을 활용하는 방법을 제안한다. 커널 밀도 추정은 비모수적 분포들에도 사용할 수 있어 국부적인 데이터 밀도 추정에 유용하다. 제안한 방법에서는 커널 밀도 추정을 수행한 후 밀도가 높은 지역에 클러스터의 초기 원형을 설정하고 원형이 설정된 영역의 밀도를 감소시키는 과정을 반복함으로써 효율적으로 초기 원형을 선택할 수 있다. 제안된 방법이 일반적으로 사용되는 무작위 초기화 방법에 비해 효율적이라는 사실은 실험 결과를 통해 확인할 수 있다.

ABSTRACT

Fuzzy C-Means (FCM) is one of the most widely used clustering algorithms and has been used in many applications successfully. However, FCM has some shortcomings and initial prototype selection is one of them. As FCM is only guaranteed to converge on a local optimum, different initial prototype results in different clustering. Therefore, much care should be given to the selection of initial prototype. In this paper, a new initialization method for FCM using kernel density estimation (KDE) is proposed to resolve the initialization problem. KDE can be used to estimate non-parametric data distribution and is useful in estimating local density. After KDE, in the proposed method, one initial point is placed at the most dense region and the density of that region is reduced. By iterating the process, initial prototype can be obtained. The initial prototype such obtained showed better result than the randomly selected one commonly used in FCM, which was demonstrated by experimental results.

키워드

군집화, Fuzzy C-Means, 초기 원형 설정, 커널 밀도 추정

Key word

Clustering, Fuzzy C-Means, Initial Prototype Selection, Kernel Density Estimation

* 정회원 : 동의대학교 영상미디어 센터
** 정회원 : 신라대학교 컴퓨터공학과 (교신저자, gbkim@silla.ac.kr)

접수일자 : 2011. 04. 05
심사완료일자 : 2011. 04. 24

Ⅰ. 서 론

군집화(clustering)는 주어진 데이터 집합 $X=\{x_1,...,x_n\}$를 K개의 균일한 부분집합으로 나누는 대표적인 비교사 학습 방법으로 다양한 분야에서 다양한 형태의 군집화 알고리듬이 사용되고 있다[1].

현재 사용되고 있는 군집화 기법들은 크게 계층적 군집화(hierarchical clustering)와 분할 기반 군집화(partitional clustering)로 나누어볼 수 있다. 계층적 군집화는 클러스터의 계층 구조를 구성하는 방식으로 하나의 클러스터에서 시작해서 연속적으로 클러스터를 나누어 가는 하향식 방법과 하나의 데이터 포인트로 구성되는 n 개의 클러스터에서 시작해서 클러스터를 뭉쳐나가는 상향식 방법이 있다. 이에 비해 분할 기반 군집화는 K개의 원형(prototype)을 설정하고 각 데이터 포이트를 가장 가까이에 위치한 원형에 할당하는 과정을 반복함으로써 K개 클러스터를 찾아내는 방식이다[1].

Fuzzy C-Means(FCM)는 대표적인 분할 기반 군집화 기법으로 1970년대 처음 소개된 이후 원형 그대로 또는 주어진 문제에 맞게 변형된 형태로 많은 문제에 성공적으로 적용되어 왔다[2][3][4][5]. 하지만 많은 FCM의 변형이 존재한다는 사실은 FCM이 모든 문제에 적합한 것은 아니라는 반증이 될 수 있다. FCM의 문제점으로는 초기 원형의 설정 문제, 가우시안 분포만을 다룰 수 있는 문제, 클러스터의 개수 설정 문제 등이 있으며 이 논문에서는 FCM이 가지는 초기 원형 설정 문제점을 살펴보고 이를 해결할 수 있는 방법을 제안한다.

초기 원형 설정을 위해 이 논문에서는 커널 밀도 추정(kernel density estimation) 기법을 이용한다. FCM은 데이터가 밀집된 지역에 클러스터를 생성하므로, 제안한 방법에서는 커널 밀도 추정을 통해 추정된 데이터 밀집 지역에 클러스터의 초기 원형을 두는 것을 기본으로 한다. 가우스 혼합 모델 등의 모수적 분포를 사용할 수도 있지만 모수적인 혼합 모델을 사용하는 경우 계산 과정은 간단해지지만 데이터가 주어진 분포를 따라야 한다는 가정으로 인해 표현력에 한계를 가진다[7]. 하지만 커널을 이용한 비모수적 방법은 데이터가 알려진 분포를 따른다는 가정을 하지 않으므로 다양한 형태의 데이터를 묘사할 수 있는 장점이 있어 확장성과 표현력에서 모수적 방법에 비해 우수하다[6][7].

제안하는 방법에서는 먼저 데이터 집합의 커널 밀도 추정을 수행한 후, 계산된 결과를 K개의 영역으로 분할하는 임계치를 동적으로 결정하고 분할된 K개 영역을 라벨링(labeling)한다. 이후 커널 밀도 추정의 결과 값 중에서 가장 높은 밀도 값을 가지는 위치를 탐색하고, 그 위치에 해당하는 라벨을 갖는 영역의 밀도 중 최소값으로 그 영역의 밀도를 조정한다. 이 때 조정된 영역의 중심은 초기 원형에 추가된다. 이러한 밀도 조정 과정은 초기 원형이 한 번 선택된 영역에서는 원형이 선택되지 않도록 하는 역할을 한다. 이 과정을 K회 반복함으로써 밀도가 높은 지역에 K개 초기 원형을 둘 수 있으며 이는 데이터 포인트를 이용한 무작위 초기화에 비해 보다 나은 해에 수렴할 가능성을 높일 수 있다. 이러한 사실은 실험 결과를 통해서 확인할 수 있다.

Ⅱ. FCM의 초기 원형 설정

FCM이 국부 최적해에 수렴하며 FCM을 이용한 군집화에는 수많은 국부 최적해가 존재한다는 것은 널리 알려진 사실이다. 이러한 국부 최적해는 초기 원형 설정 문제와 깊이 관련되어 있다. 즉, 초기 원형의 설정에 따라 수렴하는 국부 최적해가 달라지며 전역 최적해에 수렴하는 초기 원형의 설정 문제는 NP-hard임이 증명되었다[8]. 따라서 적절한 초기화를 통해 보다 나은 국부 최적해에 수렴하도록 하는 것은 중요한 연구 과제 중 하나이다. 초기화 문제를 해결하기 위해 일반적으로 무작위 초기화를 통한 군집화를 여러 번 수행한 후 그 중 최적의 결과를 선택하거나 결과들을 조합하는 등의 방법을 사용한다. 하지만 이는 연산량 요구가 높은 문제점이 있다. 이 논문에서는 FCM에서의 초기화 문제를 개선하기 위해 커널 밀도 추정을 이용한 초기 원형 선택 방법을 제안한다. 제안하는 방법은 커널 밀도 추정을 통해 국부적인 데이터의 밀도를 추정한 후 이를 바탕으로 데이터가 밀집된 지역에 초기 원형을 두는 반복 알고리듬이다. 제안하는 방법은 커널 밀도 추정을 사용함으로써 모수적 분포 추정에 비해 비정형적인 분표에서도 사용할 수 있으며 국부적인 밀도 추정이 용이한 장점이 있다. 제안하는 알고리듬은 그림 1과 같이 요약할 수 있다.

1: 데이터의 밀도 함수를 계산한다.
2: 계산된 밀도 함수가 K개 영역으로 분할되도록 이진화하고 라벨링한다.
3: k←0
3: do
4: 밀도 함수에서 밀도가 가장 높은 위치를 탐색한다.
5: 해당 위치에 대응하는 라벨링된 영역의 중점을 계산한다.
6: 중점을 초기 원형에 추가한다.
7: 해당 라벨링 영역의 밀도를 조정한다.
8: k←k+1
9: while k < K
10: return

그림 1. 커널 밀도 추정을 이용한 FCM의 초기화
Fig. 1 Initializing FCM using kernel density estimation

2.1. 커널 밀도 추정

밀도 추정에서 확률 변수 X가 IID일 때 관찰된 자료 $\{X_1, X_2, ..., X_n\}$는 모두 $1/n$ 확률을 갖게 되므로 $X=x$에서의 경험적 누적분포함수는 총 관찰치 중에서 x보다 작거나 같은 값을 갖는 자료수의 비율이 된다[6].

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \le x) \tag{1}$$

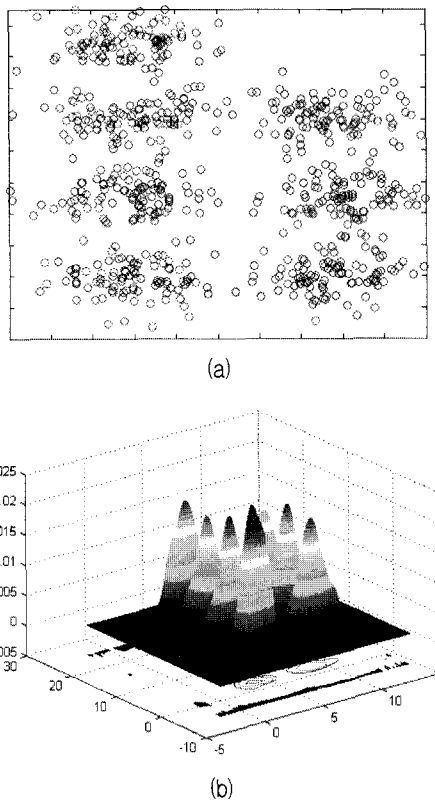
식 (1)에서 1(A) 는 A가 사실이면 1, 거짓이면 0의 값을 갖는 지시 함수를 나타낸다. 이처럼 히스토그램 방식을 이용하는 경우 계산한 추정치가 연속함수가 되지 못하는 단점 때문에 일반적으로 히스토그램의 지시함수를 연속함수 κ로 대체시킨 식(2)를 이용해 밀도 추정을 수행한다.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} \kappa\left(\frac{X_i - x}{h}\right) \tag{2}$$

식 (2)에서 h는 구간의 크기이고, n은 데이터의 개수이다. 이 논문에서는 식 (3)의 가우시안 함수를 이용하여 밀도 추정을 수행하였다[7].

$$\hat{f}'(x) = C\sum_{i=1}^{n} \exp\left(-\frac{\|X_i - x\|^2}{h}\right) \tag{3}$$

이 때 C는 정규화 상수를 나타낸다. h는 가우시안 함수의 폭으로 이 논문에서는 데이터의 분산으로 설정하였다. 그림 2는 커널 밀도 추정의 예를 보여주고 있다.



(a)

(b)

그림 2. (a) 데이터 집합과 (b) 커널을 이용하여 추정된 밀도 함수
Fig. 2 (a) Data set and (b) its estimated density function using kernels

2.2. 라벨링

커널 밀도 추정의 결과값이 주어진 K개의 영역으로 분할될 수 있는 임계치를 동적으로 결정하여 영역을 분할하고 각 영역을 라벨링한다. 영역을 분할하는 알고리듬은 그림 3과 같다. 그림 3에서 RC는 현재 임계치 값으로 이진화를 수행한 경우 찾아지는 연결 요소들의 개수를 나타낸다. 이진화를 위한 임계치는 초기에 밀도의 평균으로 시작하며 반복적으로 임계치 값을 조절함으로써 원하는 개수의 연결 요소들을 가지도록 한다.

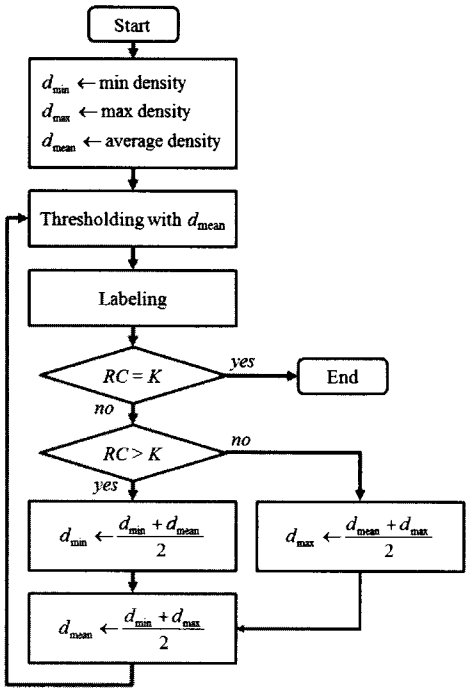


그림 3. 영역 분할 알고리듬
Fig. 3 Region split algorithm

그림 4는 그림 2에서 추정된 밀도 함수를 7개의 영역으로 나누어 라벨링한 결과를 나타내고 있다.

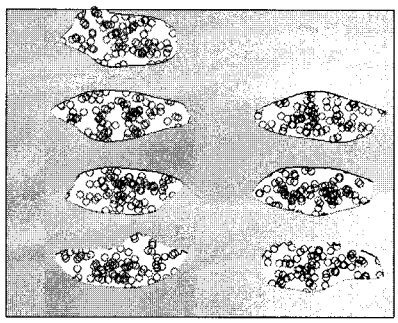
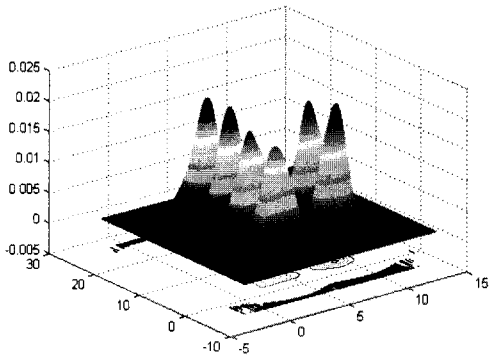


그림 4. 라벨링된 밀도 함수
Fig. 4 Labeled density function

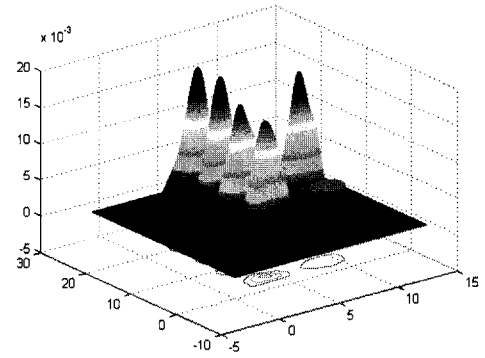
2.3. 초기 원형 설정

밀도 추정 후, 추정된 밀도 값 중 가장 큰 값을 가지는 위치의 라벨을 추출한 후, 해당 영역의 중점을 계산한다. 계산된 중점은 초기 원형에 추가된다. 이 후 해당 영역의 최저 밀도 값으로 해당 라벨 영역의 밀도를 갱신한다.

이 과정을 반복하여 주어진 클러스터의 개수 K 만큼 초기 원형을 생성한다. 그림 5는 해당 영역을 최저 밀도 값으로 대입한 영상이다. 초기 원형을 선택하는 기본 규칙 중 하나는 원형들 사이의 거리가 가능한 커야한다는 것이다. FCM은 클러스터 내의 밀집도가 높은 클러스터들을 생성하므로 초기화 과정에서 클러스터 사이의 거리를 크게 하는 것이 군집화 결과에 도움이 되는 것으로 알려져 있다[9]. 그림 5에서 보아 알 수 있듯이, 원형이 선택된 영역의 밀도를 감소시키는 것은 인접한 영역에서 다시 원형이 선택되는 것을 방지하여 원형들이 가능한 멀리 떨어지도록 해준다.



(a)



(b)

그림 5. 해당 라벨의 최저 밀도 값으로 대입
(a) 이전 (b) 이후
Fig. 5 Density substraction using the minimum density of the labeled region (a) before (b) after

Ⅲ. 실험 결과

제안된 방법의 유효성을 보이기 위해 Core i5 CPU, 4G RAM을 사용하는 컴퓨터에서 Matlab으로 각 알고리듬을 구현하고 실험하였다. 이 논문에서 비교 대상으로 삼은 알고리듬은 무작위 초기화 방법으로 이 방법은 현재에도 가장 많이 사용되는 방법 중 하나이다. 여러 정교한 초기화 방법이 존재하지만 이들 대부분이 문제의 지식을 이용하는 방법으로 일반적인 상황에서는 사용하기 어렵다. 또한 이 논문에서는 특정 데이터를 가정하지 않으므로 무작위 초기화 방법과 비교 실험하였다.

실험에 사용한 데이터는 그림 2의 데이터로 7개의 가우시안 분포로 구성되는 가우시안 혼합 모델로부터 무작위로 생성된다. 각 가우시안 분포의 평균은 μ={(3, 3), (3, 8), (3, 13), (3, 18), (8, 3), (8, 8), (8, 13)}이며 단위행렬을 공분산 행렬로 가진다. 실험은 총 100회 이루어졌으며 그 중 처음 10회에 대한 실험 결과가 표 1에 나타나 있다. 표 1에서 주어진 값은 데이터 생성에 사용된 클러스터의 중심과 군집화의 결과로 얻은 클러스터 중심 사이의 거리를 합한 것으로 식 (4)와 같이 계산된다.

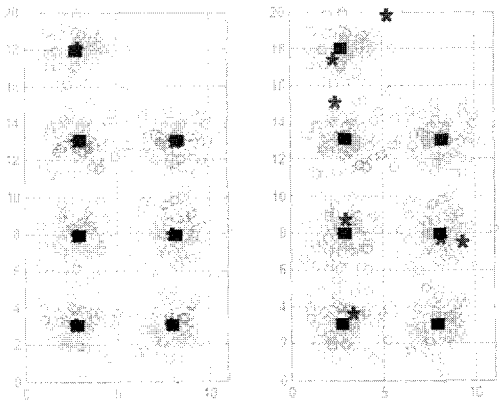
$$d=\sum_{i=1}^{7}\|\mu_i-c_i\|^2 \tag{4}$$

이 때 c_i는 클러스터링 결과로 얻어진 i번째 클러스터의 중심을 나타낸다. 표 1에서 알 수 있듯이 제안된 방법이 무작위 초기화에 비해 실제 클러스터 중심에 보다 가까운 클러스터 중심을 얻어냄을 알 수 있다. 즉, 제안한 방법이 전역 최적해에 가까운 해에 수렴할 가능성을 높여준다.

그림 6은 초기화에 따른 군집화 결과를 비교한 예이다. 그림 6에서 초록색 '*' 점은 초기 원형이고 붉은색 '■'은 데이터 생성에 사용된 실제 원형이다. 그림 6(a)에서 보아 알 수 있듯이 제안한 방법에서 초기 원형이 실제 클러스터의 중심과 상당히 유사함을 알 수 있다. 그림 6(a)에서 초기 원형은 실제 원형과 대부분 겹쳐 그림에서는 초기 원형이 명확히 나타나지 않고 있다. 이러한 초기 원형 설정 방법은 FCM이 국부 최적해에 빠질 가능성을 줄임과 동시에 수렴 속도를 높여주는 역할을 한다. 표 1에 주어진 실험에서 무작위 초기화 방법을 이용한 경우 수렴하기 위한 평균 반복 회수는 9.3회인데 반해 제안한 방법의 경우는 3.2회였다. 또한 밀도가 높은 곳에 초기 원형을 설정함으로써 군집화 과정에서 원형의 변화가 거의 없고 이동 거리가 매우 적음을 확인할 수 있다.

표 1. 실험 결과
Table. 1 Experimental results

	(a) 제안된 방법	(b) 무작위 방법
1	0.6954	13.612
2	0.7371	18.993
3	0.5892	17.788
4	0.5383	24.928
5	1.1993	17.2055
6	0.5694	22.0083
7	0.7555	31.4492
8	0.8311	12.0834
9	0.6719	21.7241
10	0.5858	15.9253



(a) (b)

그림 6. (a) 제안하는 방법과 (b) 무작위 초기화 방법의 결과 비교
Fig. 6 Comparison between the proposed method and the random initialization method

Ⅳ. 결 론

FCM은 대표적인 분할 기반 군집화 기법으로 지금까지도 널리 사용되는 대표적인 군집화 알고리듬 기법 중

의 하나이다. 하지만 FCM은 여러 가지 해결되지 못한 문제점들을 가지고 있으며 이 논문에서는 그 문제점들 중 하나인 초기 원형 설정 문제를 개선하는 방법을 제안하였다. 제안한 방법에서는 커널 밀도 추정을 활용하여 데이터가 밀집된 지역에 클러스터의 초기 원형을 둠으로써 기존에 사용되는 무작위 초기화 방법에 비해 나은 결과를 얻을 수 있었다. 하지만 제안한 방법 역시 개선의 여지는 남아 있다. 커널 밀도 추정 방법이 국부적인 밀도 추정에 유용한 것은 사실이지만 데이터의 수의 제곱에 비례하는 연산을 요구하는 단점이 있다. 따라서 연산량이 적으면서도 커널 밀도 추정과 유사한 성능을 보이는 함수를 개발하는 것이 대용량의 데이터를 처리하기에 필요하며 이는 현재 연구 중에 있다.

참고문헌

[1] R. Xu and D. Wunsch, *Clustering*, Wiley-IEEE Press, 2008.

[2] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338-353, 1965.

[3] E. H. Ruspini, "A new approach to clustering," *Information and Control*, vol. 16, pp. 22-32, 1969.

[4] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters," *Journal of Cybernetics*, pp. 32-57, 1974.

[5] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer, 1981.

[6] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2001.

[7] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Annals of Statistics*, vol. 38, no. 5, pp. 2916-2957, 2010.

[8] M. Garey, D. Johnson, and H. Witsenhausen, "The complexity of the generalized Lloyd-Max problem," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 255-256, 1982.

[9] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-847, 1991.

저자소개

허경용(Gyeongyong Heo)

1996년 8월 : 연세대학교 본대학원 전자공학과 (공학석사)
2009년 12월: Dept. of Computer and Information Science and Engineering, University of Florida (공학박사)

※관심분야 : Machine Learning, Pattern Recognition, Image Processing

김광백(Kwang-Baek Kim)

1999년 : 부산대학교 전자계산학과 (이학박사)
1997년~현재 : 신라대학교 컴퓨터공학과 교수
2005년~현재 : 한국멀티미디어학회 이사
2005년~현재 : 한국해양정보통신학회 학술상임이사

※관심분야 : Image Processing, Fuzzy Logic, Neural Networks, Medical Imaging and Biomedical System, Support Vector Machines