한국컴퓨터정보학회 동계학술대회 논문집 제19권 제1호 (2011. 1)

커널 밀도 추정을 이용한 Fuzzy C-means의 초기 원형 설정

조현학°, 허경용**, 김광백*
°*신라대학교 컴퓨터정보공학부
**동의대학교 영상미디어센터
E-mail: darkruby1004@naver.com, Gyeongyong.Heo@gmail.com, gbkim@silla.ac.kr

Initial Prototype Selection in Fuzzy C-Means Using Kernel Density Estimation

Hyun-Hak Cho°, Gyeongyong Heo**, Kwang-Beak Kim*
°*Division of Computer and Information Engineering, Silla University
**Visual Media Center, Dong-Eui University

● 요 약 ●

Fuzzy C-Means (FCM) 알고리듬은 가장 널리 사용되는 군집화 알고리듬 중 하나로 다양한 응용 분야에서 사용되고 있다. 하지만 FCM은 여러 가지 문제점을 가지고 있으며 초기 원형 설정이 그 중 하나이다. FCM은 국부 최적해에 수렴하므로 초기 원형 설정에 따라 클러스터링 결과가 달라진다. 이 논문에서는 이러한 FCM의 초기 원형 설정 문제를 개선하기 위하여 커널 밀도 추정 (kernel density estimation) 기법을 활용하는 방법을 제안한다. 제안한 방법에서는 먼저 커널 밀도 추정을 수행한 후 밀도가 높은 지역에 클러스터의 초기 원형을 설정하고 원형이 설정된 영역의 밀도를 감소시키는 과정을 반복함으로써 효율적으로 초기 원형을 설정할 수 있다. 제안된 방법이 일반적으로 사용되는 무작위 초기화 방법에 비해 효율적이라는 사실은 실험 결과를 통해 확인할 수 있다.

키워드: Fuzzy c-means (Fuzzy c-means), 초기 원형 설정 (Initial Prototype Selection), 커널 밀도 추정 (Kernel Density Estimation)

I. 서 론

클러스터링은 주어진 데이터 집합 $X=\{x_1,...,x_n\}$를 K 개의 균일한 부분집합으로 나누는 대표적인 비교사 학습 방법으로 다양한 분야에서 다양한 형태의 클러스터링 알고리듬이 사용되고 있다[1-3].

현재 사용되고 있는 클러스터링 기법들은 크게 계층적 클러스터링 (hierarchical clustering)과 분할 기반 클러스터링 (partitional clustering)으로 나누어볼 수 있다. 계층적 클러스터링은 클러스터의 계층 구조를 구성하는 방식으로 하나의 클러스터에서 시작해서 연속적으로 클러스터를 나누어 가는 하향식 방법과 하나의 데이터 포인트로 구성되는 n 개의 클러스터에서 시작해서 클러스터를 포함해가는 상향식 방법이 있다. 이에 비해 분할 기반 클러스터링은 K 개의 원형(prototype)을 설정하고 각 데이터 포이트를 가장 가까이에 위치한 원형에 할당하는 과정을 반복함으로써 K 개 원형을 찾아내는 방식이다[4-6].

Fuzzy C-Means(FCM)는 대표적인 분할 기반 클러스터링 기법으로 1970년대 처음 소개된 이후 원형 그대로 또는 주어진 문제에 맞게 변형된 형태로 많은 문제에 성공적으로 적용되어 왔다. 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. 기본적으로 FCM은 데이터가 밀집된 지역에 클러스터를 생성하는 것이 자연스러우므로 커널 밀도 추정을 통해 추정된 밀집 지역에 클러스터의 초기 원형을 두는 것을 기본으로 한다. 가우스 혼합 모델 등을 사용하여 유사한 효과를 얻을 수 있지만 모수적인 혼합 모델을 사용하는 경우 계산 과정은 간단해질 수는 있지만 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 기존 방법에 우수하다[8].

제안된 방법에서는 먼저 데이터 집합의 커널 밀도 추정을 수행한 후, 계산된 결과를 K 개의 영역으로 분할하는 임계치를 동적으로 결정하고 각 영역을 라벨링(labeling)한다. 이후 커널 밀도 추정의 결과 값 중에서 가장 높은 밀도 값을 가지는 위치를 탐색

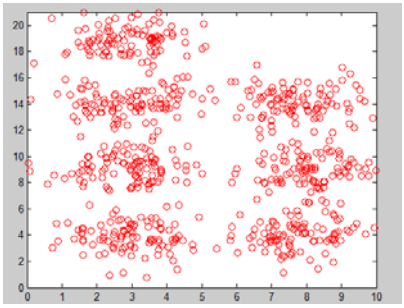
하고, 그 위치에 해당하는 라벨을 갖는 영역의 밀도 중 최소값으로 그 영역의 밀도를 조정한다. 이 때 조정된 영역의 중심은 초기 원형으로 추가된다. 이러한 조정 과정은 초기 원형이 한 번 선택된 영역에서는 다시 원형이 선택되지 않도록 하는 역할을 한다. 이 과정을 K회 반복함으로써 밀도가 높은 지역에 K개 초기 원형을 둘 수 있으며 이는 데이터 포인트를 이용한 무작위 초기화에 보다 나은 해에 수렴할 가능성을 높일 수 있다. 이러한 사실은 실험 결과를 통해서 확인할 수 있다.

II. 제안하는 초기 원형 설정 방법

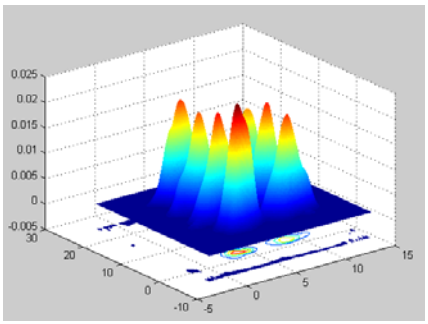
FCM이 국부 최적해에 수렴한다는 것은 널리 알려진 사실이며 FCM을 이용한 클러스터링에는 수많은 국부 최적해가 존재한다는 것도 역시 널리 알려져 있다. 이러한 국부 최적해는 또한 초기 원형 설정 문제와 깊이 관련되어 있다. 즉, 초기 원형의 설정에 따라 수렴하는 국부 최적해가 달라지며 전역 최적해에 수렴하는 초기 원형의 설정 문제는 NP-hard임이 증명되었으므로 적절한 초기화를 통해 보다 나은 국부 최적해에 수렴하도록 하는 것은 중요한 연구 과제 중 하나이다. 이러한 초기화 문제를 해결하기 위해 일반적으로 무작위 초기화를 통한 클러스터링을 여러번 수행한 후 그 중 최적의 결과를 선택하거나 결과들을 조합하는 등의 방법을 사용한다. 하지만 이는 추가적인 연산을 필요로하는 문제점이 있다.

이 논문에서는 이러한 초기화 문제를 개선하기 위해 커널 밀도 추정을 이용한 초기 원형 선택 방법을 제안한다. 제안하는 알고리듬은 다음과 같다.

1: 데이터 집합을 양수화 한다.
2: 데이터의 커널 밀도 함수를 계산한다.
3: 계산된 커널 밀도 함수가 K개 영역으로 분할되도 록 이진화하고 라벨링한다.
4: **do**
5: 커널 밀도 함수에서 밀도가 가장 높은 위치를 탐색한다.
6: 해당 위치에 대응하는 라벨링된 영역의 중점을 계산한다.
7: 해당 라벨링 영역을 최소 밀도로 조정한다.
8: K = K + 1
9: **while K 〈 주어진 원형의 개수**
10:**return**



(a)



(b)

그림 1. (a) 데이터 집합과 (b) 커널 밀도 추정
Fig. 1. (a) Data set and (b) its estimated density using kernels

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. 즉 $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \le x)$ 이며 여기서 1(A) 는 A가 사실이면 1의 값을 갖고 A가 거짓이면 0의 값을 갖는 지시 함수를 나타낸다. 이처럼 히스토그램 방식을 이용하는 경우 계산한 추정치가 연속함수가 되지 못하는 단점 때문에 히스토그램의 지시함수를 연속함수 κ로 대체시킨 식(1)을 이용해 커널 밀도 추정을 수행한다.

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

식 (1)에서 h는 구간의 크기이고, n은 데이터의 개수이다. 그림 1은 커널 밀도 추정의 예를 보여주고 있다.

2. 라벨링 (Labeling)

커널 밀도 추정의 결과값이 주어진 K 개의 영역으로 분할될 수 있는 임계치를 동적으로 결정하여 영역을 분할하고 각 영역을 라벨링한다. 영역을 분할하는 알고리듬은 다음과 같다.

```
1: mean = Kernel Density Estimator mean value
2: max  = Kernel Density Estimator max value
3: min  = Kernel Density Estimator min value
4: before_k = k
5: do
6: Threshold(Kernel Density Estimator, mean)
7: labeling
8: if label_count 〉 k then
9:     if label_count 〈 before_k then
10:        min = mean
11:        mean = (max + min) / 2
12:    else
13:        max = mean
14:        mean = (max + min) / 2
15:    end
16: else
17:    if label_count 〈 before_k then
18:        min = mean;
19:        mean = (max + min) / 2
20:    else
21:        max = mean
22:        mean = (max + min ) /2
23:    end
24: end
25: before_k = label_count
26: while K ~= label_count
```

그림 2는 그림 1에서 추정된 밀도 함수를 7개의 영역으로 나누어 라벨링한 결과를 나타내고 있다.

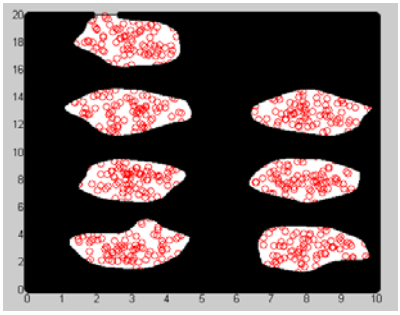
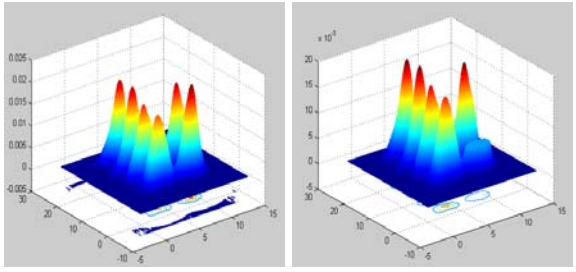


그림 2. 라벨링된 밀도 함수
Fig. 2. Labeled density function

3. 초기 원형 설정

커널 밀도 추정의 결과 값 중 가장 높은 밀도 값을 가지는 위치의 라벨을 추출한 후, 해당 영역의 중점을 계산한다. 측정한 이후 중점을 해당 라벨의 초기 원형으로 추가한다. 이 후 해당 영역의 최저 밀도 값을 구한 후, 해당 라벨 영역에 대입한다. 이 과정을 반복하여 주어진 클러스터의 개수 K 만큼 초기 원형을 측정한다. 그림 3은 해당 영역을 최저 밀도값으로 대입한 영상이다.



(a) before (b) after

그림 3. 해당 라벨 최저 밀도 값으로 대입 영상
Fig. 3. Density substraction using the minimum density of the labeled region

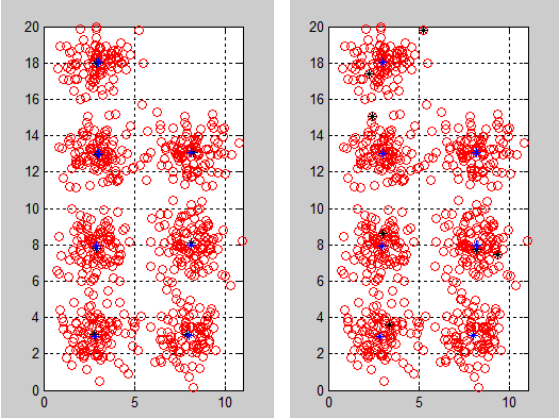
III. 실험 및 결과

제안된 방법의 유효성을 보이기 위해 이 논문에서는 가장 많이 사용되는 무작위 초기화 방법과 비교하였다. 실험에 사용한 데이터는 그림 1의 데이터이며 표 1은 실험 결과를 요약한 것이다. 표 1에서 주어진 값은 데이터 생성에 사용된 클러스터의 중심과 클러스터링 결과값으로 얻은 클러스터 중심 사이의 거리를 합한 것이다. 표 1에서 알 수 있듯이 제안된 방법이 무작위 초기화에 비해 실제 클러스터 중심에 보다 가까운 클러스터 중심을 얻어냄을 알 수 있다. 또한 밀도가 높은 곳에 초기 원형을 설정함으로써 클러스터링 과정에서 원형의 변화가 거의 없고 이동 거리가 매우 적음을 확인할 수 있다[9-10].

표 1. 실험 결과
Table 1. Experimental results

	(a)제안된 방법	(b)FCM
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

그림 4는 초기화에 따른 클러스터링 결과를 비교한 예이다. 그림 4에서 검은 '*' 점은 초기 원형이고 파란색의 '*'은 FCM 수행 후에 결과 원형이다. 그림 5(a)의 초기 원형이 그림 5(b)의 초기 원형과 비교한 결과, 제안한 방법이 7개의 그룹에 초기 원형 설정이 상당히 효율적으로 분할된 것을 확인할 수 있다.



(a) 제안하는 방법 (b) 무작위 초기화 방법

그림 4. 제안된 방법과 무작위 초기화의 결과 비교
Fig. 4. Comparison between the proposed method and random initialization method

Ⅳ. 결 론

FCM은 오랜 역사를 지니고 있고, 지금까지도 널리 사용되는 대표적인 클러스터링 알고리즘 기법 중의 하나이지만 여러 가지 해결되지 못한 문제들이 있다. 이 논문에서는 그 문제점들 중에서 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 있었으며 이는 실험 결과를 통해 확인할 수 있다.

참고문헌

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

[2] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, pp. 281~297, 1967.

[3] L. A. Zadeh, “Fuzzy sets,” Information and Control Vol. 8, No. 3, pp. 338~353, June 1965.

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

[5] H. Frigui and R. Krishnapuram, “Clustering by competitive agglomeration,” Pattern Recognition, Vol. 30, No. 7, pp. 1109~1119, July 1997.

[6] G. Heo and Y.W. Woo, “Extensions of X-means with Efficient Learning the Number of Clusters,” The Journal of the Korean Institute of Maritime Information and Communication Sciences, Vol. 12, No. 4, pp. 772~780, Apr. 2008.

[7] E. H. Ruspini, “A new approach to clustering,” Information and Control, Vol. 15, No. 1, pp. 22~32, July 1969.

[8] J. C. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters,” Cybernetics and Systems, Vol. 3, No. 3, pp. 32~57, July 1974.

[9] G. Heo and P. Gader, “Learning the Number of Gaussian Components Using Hypothesis Test,” Proceedings of the 2009 International Joint Conference on Neural Networks, pp. 1206~1212, 2009.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” Journal of the Royal Statistical Society, Series B, Vol. 39, No. 1, pp. 1~38, Jan. 1977.