밀도 함수를 이용한 근사적 퍼지 클러스터링

Approximate fuzzy clustering based on a density function

손 세 호, 권 순 학, 최 윤 혁
Seo H. Son, Soon H. Kwon and Y. H. Choi
경북 경산시 대동 214-1
영남대학교 공과대학 전자정보공학부
Tel: 053-810-3514,1528
Fax: 053-813-8230
E-mail: m0040306@chunma.yeungnam.ac.kr

ABSTRACT : We introduce an approximate fuzzy clustering method, which is simple but computationally efficient, based on density functions in this paper. The density functions are defined by the number of data within the predetermined interval. Numerical examples are presented to show the validity of the proposed clustering method.

1. 서론

우리가 일반적으로 사용하는 퍼지 클러스터링의 방법에는 Beztek의 FCM (Fuzzy C-Mean)[1], FCM이 갖는 단점을 보완하기 위해 가능성(possibility)을 이용한 방법[2] 및 이들 두 방법의 결합에 의한 혼용 방법[3], 그리고 Yager와 Filev의 산 클러스터링 방법(Mountain method)[4] 등등이 있다. 이러한 방법들은 초기에 주어진 데이터를 바탕으로 클러스터링을 수행하기 때문에 초기 값의 설정이 중요하고 계산 과정이 복잡하여 계산 시간이 많이 소요되며 이해도도 떨어지는 단점을 지니고 있다. 이러한 단점을 보완하기 위해 Linkens[5] 등은 초기에 주어진 데이터에 대하여 신경회로망을 이용하여 근사적 클러스터를 구한 후, 이를 기존의 퍼지 클러스터링 알고리즘에 적용하는 계층적 퍼지 클러스터링 방법을 제시하였다. 그러나, 이 방법은 근사적 클러스터를 구함에 있어 필요한 변수 값의 설정에 따라 클러스터링 결과가 크게 변하는 단점을 지니고 있으며 또한 상당히 복잡한 구조를 지닌다는 단점을 지니고 있다.

본 논문에서는 기존의 클러스터링 알고리즘이 갖는 단점을 보완하기 위하여 간단하고 계산 시간이 기존의 퍼지 클러스터링 알고리즘에 비하여 현저히 적으면서도 성능 저하가 비교적 적은 계층적 구조를 갖는 근사적 퍼지 클러스터링 알고리즘을 제시하고자 한다. 제안된 계층적 구조를 갖는 근사적 퍼지 클러스터링 알고리즘의 하부 계층에서는 주어진 데이터의 분포를 이용하여 밀도 함수를 생성하고 이를 바탕으로 근사적 클러스터를 생성한다. 상부 계층은 일반적인 FCM으로 구성되어 있으며, 하부 계층에서 생성된 근사적 클러스터를 이용하여 퍼지 클러스터링을 수행한다.

2. 밀도 함수를 이용한 근사적 퍼지 클러스터링

이 절에서는, 본 논문에서 제안하는 밀도 함수를 이용한 근사적 퍼지 클러스터링 방법의 도입 배경과 클러스터링 방법에 대해 다루고자 한다. 일반적으로, 클러스터 중심 주변에는 많은 양의 자료들이 분포하고 있음을 알 수 있다. 이를 바탕으로 자료가 주어지는 공간을 여러개의 부분 공간으로 분할하여 부분 공간속에 속하는 자료의 개수를 자료의 밀도(density)로 정의한다면 클러스터 중심의 밀도가 다른 부분 공간의 밀도보다 높음을 알 수 있다[6]. 즉, 밀도가 높으면 높을수록 클러스터의 중심이 될 가능성이 높다고 할 수 있다. 이와 같은 특성을 이용하여 근사적인 클러스터의 수를 결정하

는 것이 본 논문에서 보이고자 하는 클러스터링 방법이다.

이러한 밀도 함수를 이용한 근사적 퍼지 클러스터의 중심을 구하는 과정에서 중요한 문제점의 하나는 주어진 자료 공간의 분할과 분할된 자료들로부터 밀도를 계산하는 것이다. 본 논문에서 제시하는 클러스터링 방법은 Yager의 산 클러스터링 방법에서 사용한 자료 공간의 분할이나 밀도 함수를 이용한 점에서 유사하다고 할 수 있으나 산 클러스터링 방법은 자료 공간을 임의로 분할하고 분할하는 선들의 교점에서 잠재적인 클러스터를 찾아 복잡하게 정의된 밀도 함수를 계산하여 밀도를 얻는다. 그러나, 본 논문에서 제시한 방법은 자료의 특성에 따라 분할된 영역 속에 포함되는 자료의 개수를 측정한 밀도를 이용하여 근사적인 클러스터의 중심을 찾는다는 것이 Yager의 산 클러스터링 방법과 본 논문에서 제시한 방법의 큰 차이점이라 할 수 있다.

이러한 근사적 클러스터를 구하는 과정은 다음의 5단계 과정을 통해 이루어지며, 고려 대상 자료의 차원을 알고리즘 설명의 편의상 2차원 공간에 한정하기로 한다 .

(1) 주어진 모든 자료 $a_j=(a_j^1, a_j^2)$ $j=1,2,\ldots,N$를 각 좌표계 a^1축, a^2축 상에 투영시킨다.

(2) 각 점에서의 이웃하는 점까지의 최소 거리를 구한 후 그 중 최대값을 구한다.

$$\Delta d^i = \max_{k\in j}\left(\min_{l\in j}(\|a_k^i - a_l^i\|)\right)$$

(단, $i=1,2$, $j=1,\ldots,N$, $k\neq l$)

(3) (2)에서 얻어진 값 Δd^i에 적절한 상수 M을 곱한 값 $M\times\Delta d^i=\Delta d^i_M$을 이용하여 각 좌표계를 다음과 같이 분할한다.

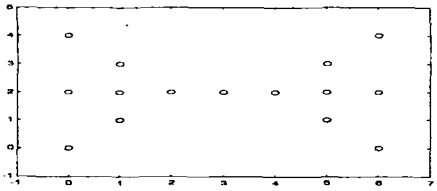
$a^i_{min}+n\times\Delta d^i_M \le a^i < a^i_{min}+(n+1)\times\Delta d^i_M$,

(4) 분할된 각 구간의 밀도를 구한다.

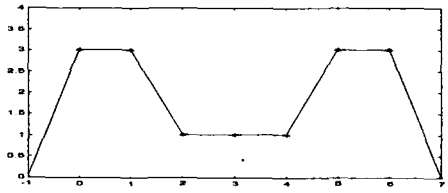
(5) (4)에서 구한 각 구간의 밀도를 이용하여 근사적 클러스터의 수 및 클러스터 중심을 구한다.

(6) (5)에서 구해진 클러스터 중심을 FCM의 초기치로 설정하여 FCM을 수행한다.

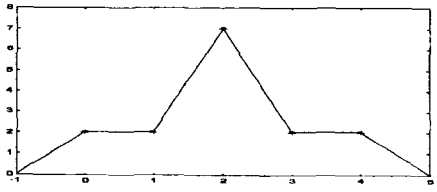
위에서 제시한 알고리즘을 일반적으로 2개의 클러스터를 가지는 나비형 자료 그림 1.(a)를 이용하여 설명하기로 한다. 먼저, 주어진 자료를 X축과 Y축에 대해 투영한 후 각 좌표계에 대한 이산 구간을 구하여 분할한다. 분할된 영역에서의 밀도를 계산하여 그림으로 나타내면 그림 1.(b) 및 (c)와 같다.



(a) 나비형 자료



(b) X축에 대한 각 구간에서의 밀도 함수



(c) Y축에 대한 각 구간에서의 밀도 함수

그림 1. 밀도 함수를 이용한 클러스터링

그림 1에서 보면 X축에서는 2개의 클러스터가 존재하고 Y축에서는 1개의 클러스터가 존재한다. 이 두 결과를 종합해 보면 그림 1의 나비형 자료는 2개의 클러스터를 가짐을 알 수 있다. 그림 1의 자료는 나비 모양의 규칙적인 형태를 갖지만, 일반적으로 자료들의 분포는 규칙적이거나 명확하지 않다. 이러한 불규칙적인 자료에서 가장 중요한 문제가 되는 것은 적절한 이산구간을 찾는 것이다. 이산 구간이 작으면 밀도 함수의 그래프는 많은 산 모양을 가지며 이산 구간이 커지면 그 수는 감소하며 결국에는 단 하나의 산 모양만을 형성한다. 즉, 단

하나의 클러스터만을 가지는 것처럼 보인다. 이러한 문제를 해결하기 위해 이산 구간을 $M \times \Delta d^i$로 설정하여 다음의 모의 실험을 통해 적절한 이산 구간을 구하는 방법을 보이고자 한다.

3. 시뮬레이션 결과 및 검토

앞에서 본 논문에서 제시한 알고리즘이 규칙적인 자료에 대해 만족함을 보였다. 이 절에서는 불규칙적인 자료-임의의 다섯 점 (2.72 , 5.62), (2.95,3.09), (5.02,5.13), (7.04,3.07) 및 (7.18,5.41) 주위에 분포한 500개의 랜덤 자료-에 대한 알고리즘의 타당성을 보이기 위해 모의 실험을 하였다. 주어진 자료의 특성을 가장 잘 나타내는 이산 구간을 설정하기 위해 이산 구간을 $M \times \Delta d^i (i=1,2)$로 선택하여 모의 실험하였다.

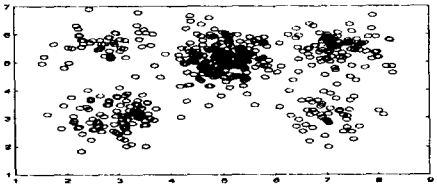
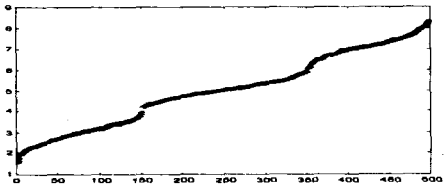
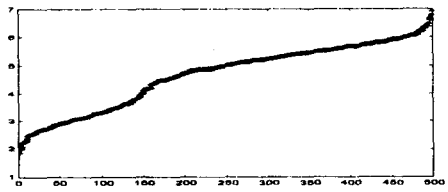


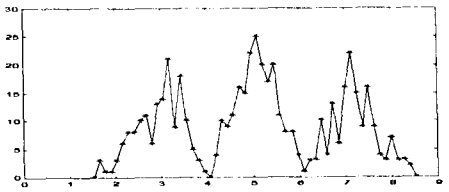
그림 2. 500개의 랜덤 자료



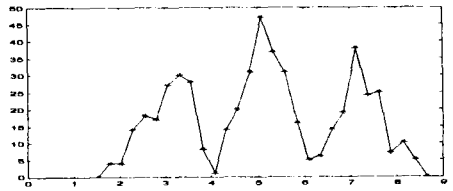
(a) 정렬된 x좌표의 분포도
※(X축=자료의 개수, Y축=x좌표)



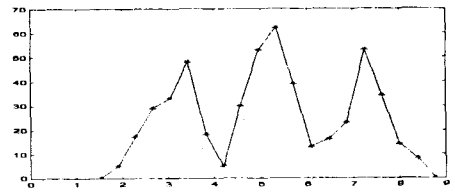
(b) 정렬된 y좌표의 분포도
※(X축=자료의 개수, Y축=y좌표)
그림 3. 정렬된 x, y좌표의 분포도



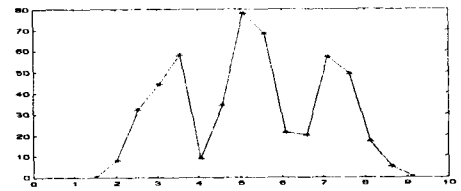
(a) M=1일 때의 밀도 함수 (X좌표)



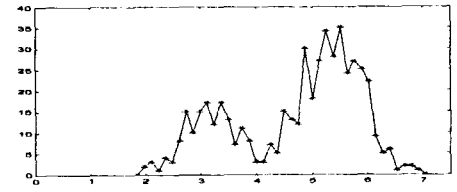
(b) M=2일 때의 밀도 함수 (X좌표)



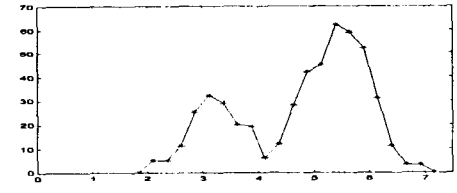
(c) M=3일 때의 밀도 함수 (X좌표)



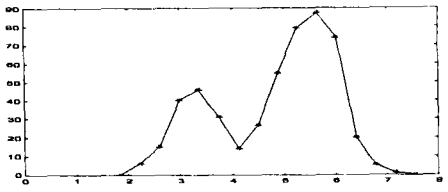
(d) M=4일 때의 밀도 함수 (X좌표)



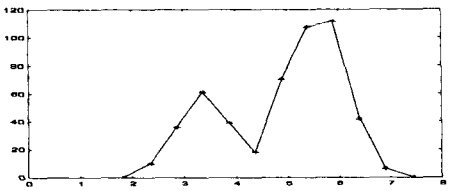
(e) M=1일 때의 밀도 함수 (Y좌표)



(f) M=2일 때의 밀도 함수 (Y좌표)



(g) M=3일 때의 밀도 함수 (Y좌표)



(h) M=4일 때의 밀도 함수 (Y좌표)

그림 4. 모의 실험 결과

※(X축=자료의 구간, Y축=자료의 개수)

그림 3의 정렬된 자료들의 분포도를 보면 자료들의 밀집 정도를 알 수 있다. 바꾸어 말해, 아주 근접한 점들의 기울기는 0에 가까운 값을 가진다. 즉, 기울기가 완만한 구간이 높은 밀도를 가지므로 클러스터 중심이 존재한다고 할 수 있다. 이런 사실을 이용하여 그림 3.(a)와 (b)를 보면 각각의 X축, Y축에 대해 3개, 2개의 근사적인 클러스터를 가짐을 알 수 있다.

그림 4.(c)와 (f)에서 M=3과 2일 때 각각의 그래프가 산 모양을 가지기 시작한다. 이산 구간이 커지면 보다 확실한 산 모양을 얻을 수 있지만 클러스터 중심의 정확성이 낮아지므로 그림 4.(c)와 (f)를 이용하여 근사적인 클러스터의 중심을 얻는 것이 바람직하다.

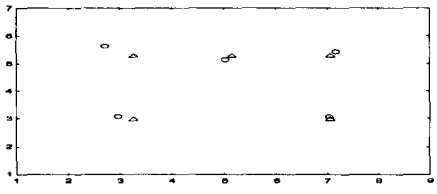


그림 5. FCM과 제안된 알고리즘의 모의 실험 결과 비교

(FCM의 중심:O, 근사적 클러스터의 중심:△)

그림 5에서 보면 그림 4.(c)와 (f)에서 얻은 밀도를 이용하여 구한 근사적 퍼지 클러스터 중심은 그림 2에 나타낸 자료의 특성과 거의 일치함을 알 수 있다.

4. 결 론

본 논문에서 제시한 밀도 함수를 이용한 퍼지 클러스터링 방법의 타당성과 앞에서 언급한 다른 클러스터링 방법보다 여러 가지 장점이 있음을 모의 실험을 통해 보였다. 향후 과제로는 모의 실험에서 보인 클러스터의 개수와 이산 구간 사이에 존재하는 관계에 대한 연구이다. 위의 관계를 알 수 있다면 본 논문에서 제시한 알고리즘은 보다 넓은 분야에 사용 될 것으로 기대된다.

5. 참고 문헌

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

[2] R. Krishnappuram and J.M. Keller, "A Possibilistic Approach to Clustering," IEEE Trans. Fuzzy Syst., vol.1, no.2, pp.98-110, 1993.

[3] N.R. Pal, K. Pal and J.C. Bezdek, "A Mixed C-Means Clustering Model," in Proc. FUZZ-IEEE'97, pp. 11-21, 1997.

[4] R.R. Yager and D.P. Filev, Essentials of fuzzy modeling and control, John Wiley & Sons, Inc., New York, 1994.

[5] D. A. Linkens and Min-You Chen, "Hierarchical Fuzzy Clustering Based on Self-organising Network," in proc. Fuzzy-IEEE, pp. 1406-1410, 1998.

[6] 권순학, 정혜천, 이석규, "밀도 함수를 이용한 퍼지 클러스터링," '99 제2회 ICASE 대구·경북지부 학술 발표회 논문집, pp. 1-4, 1999.