한국 퍼지 및 지능시스템학회 논문지 2000, Vol. 10, No. 4 (2000)285~292

밀도함수를 이용한 근사적 퍼지 클러스터링

Approximate Fuzzy Clustering Based on Density Functions

권순학 · 손세호

Soon H. Kwon and Seo H. Son

영남대학교 전자정보공학부

요 약

자료 분석 과정을 살펴 보면 1) 자료가 갖는 경향 평가, 2) 클러스터 분석, 3) 클러스터의 타당성 조사라는 과정을 거쳐 이루어진다. 이 분석법은 2) 및 3) 단계의 반복 수행으로 인하여 많은 계산 시간이 소요되므로 비효율적인 방법이라 할 수 있다. 본 논문에서는, 이와 같은 단점을 보완하기 위하여 자료가 갖는 개략적 특성을 파악하여 자료 속에 존재하는 클러스터의 근사적 개수 및 중심을 정한 후, 이 정보를 기존의 일반적인 퍼지 클러스터링 알고리즘에 입력하여 클러스터링을 수행하는 밀도함수를 이용한 계층적 구조의 근사적 클러스터링 알고리즘을 제안하고, 예제를 통하여 제안된 알고리즘의 타당성을 보인다.

ABSTRACT

In general, exploratory data analysis consists of three processes: i) assessment of clustering tendency, ii) cluster analysis, and iii) cluster validation. This analysis method requiring a number of iterations of step ii) and iii) to converge is computationally inefficient. In this paper, we propose a density function-based approximate fuzzy clustering method with a hierarchical structure which consists of two phases: Phase I is a features(i.e., number of clusters and cluster centers) extraction process based on the tendency assessment of a given data and Phase II is a standard FCM with the cluster centers initialized by the results of the Phase I. Numerical examples are presented to show the validity of the proposed clustering method.

1. 서 론

일반적으로 자료의 분석에 있어서 수행되어지는 과정을 요약하면 다음과 같이 3 단계를 거쳐 이루어진다[1,12,16-18]. 첫 번째 단계는 자료가 갖는 구조에 대한 철저한 조사 없이 자료 내부에 어떤 구조가 존재하는지를 평가하는 것으로 자료의 개략적 특성을 파악하는 과정이라 할 수 있다. 만일, 이 과정에서 자료 내부에 어떠한 구조도 존재하지 않을 것이라고 판단된다면 다음 2 단계를 수행할 필요가 없으므로 자료 분석은 여기서 끝이 나게 된다. 두 번째 단계는 클러스터 분석을 수행하는 과정으로, 자료가 갖는 성격 또는 알고리즘의 편의성에 따라 퍼지 이론에 근거한 알고리즘[2-4], 확률에 근거한 알고리즘[5] 등등이 개발되어 사용되어지고 있다.

퍼지 클러스터링 알고리즘 중에서 가장 널리 쓰이고 있는 방법은 Bezdek의 FCM (Fuzzy c-means)[2]이라 할 수 있다. 이 FCM은 주어진 자료 $X=\{x_1, x_2, \cdots, x_n\}$에 대하여 클러스터의 수를 $c(1<c<n)$로 가정하고 $n\times c$의 퍼지 분할 행렬을 반복적 연산을 통하여 생성하는 알고리즘이다. 이때 퍼지 분할 행렬에 대한 초기 값의 선정은 알고리즘의 수렴 속도에 직접적으로 영향을 미치기 때문에 FCM의 성능을 좌우하는 중요한 요소라 할 수 있다. 이와 같이 FCM이 주어진 각각의 자료가 c개의 클러스터 각각에 어느 정도 속하는지를 나타내는 퍼지 분할 행렬을 구하는 알고리즘인 것에 반하여 가능성을 이용한 방법 클러스터링(Possibilistic c-means: PCM)[3]은 주어진 각각의 자료가 어떤 클러스터에 속할 가능성, 즉, 고유성(Typicality)을 찾는 알고리즘이라 할 수 있으며, 이 방법 또한 퍼지 소속도 함수에 대한 초기값의 설정이 연산 속도 및 클러스터링 결과에 지대한 영향을 미친다고 알려져 있다. 이 이외에도 이들 두 방법의 결합에 의한 혼용 방법[4], 그리고 Yager와 Filev의 산 클러스터링 방법(Mountain method: MM)[6] 등등이 있다. MM은 자료가 정의되는 공간을 적절한 크기로 분할하고, 분할된 공간들의 교점을 클러스터의 중심이 위치할 점의 후보로 정하고, 이들 후보 중에서 미리 설정된 함수(Mountain function)값을 최적화시키는 점

이 논문은 2000학년도 영남대학교 학술연구조성비 지원에 의한 것임.

들을 클러스터 중심으로 설정하는 알고리즘이다. 이 MM 또한 초기에 자료 공간을 어떻게 분할하느냐에 따라 클러스터링 결과가 크게 달라지는 단점을 지니고 있다.

이러한 알고리즘들의 특징은 주어진 모든 데이터를 바탕으로 퍼지 클러스터링을 수행하기 때문에 계산 시간이 많이 소요되는 단점을 지니고 있다. 이러한 단점을 보완하기 위해 Cannon등[7]은 알고리즘 수행 과정에서 요구되는 자료를 정수화하여 FCM을 구성하는 방식을 제안하였으며, Linkens[8] 등은 초기에 주어진 데이터에 대하여 신경회로망을 이용하여 자료들의 합병을 통한 근사적 클러스터를 구한 후, 이를 기존의 퍼지 클러스터링 알고리즘에 적용하는 계층적 퍼지 클러스터링 방법을 제시하였으며, Cheng[9] 등도 Linkens[8]의 알고리즘 및 Kalviainen[10]의 알고리즘과 유사 개념의 알고리즘을 제시하였다. 그러나, 이들 방법 또한 근사적 클러스터를 구함에 있어 필요한 변수 값의 설정에 따라 클러스터링 결과가 크게 변화하며 또한 상당히 복잡한 구조를 지닌다는 단점을 지니고 있다.

자료 분석의 마지막 단계는 c개로 분할된 클러스터에 대한 타당성을 검토하는 과정, 즉, c개로 가정하여 얻어진 퍼지 클러스터에 관한 정보를 바탕으로 자료 속에 존재하는 클러스터의 수 c^*에 근접하는 c 값을 발견하는 과정[14]이다. 이를 위한 지표로는 Dunn의 분할 지표[11], Xie와 Beni의 지표[13], Xie와 Beni 지표가 갖는 단점, 즉, 지표 값이 클러스터의 수가 증가함에 따라 단조 감소하는 현상을 보완하여 지표의 성능을 향상시킨 Kwon의 지표[15] 등등이 사용되어지고 있다.

위에서 살펴 본 자료 분석 과정을 요약하면 1) 자료가 갖는 경향 평가, 2) 클러스터 분석, 3) 클러스터의 타당성 조사가 된다. 여기서, 2) 및 3) 단계를 반복 수행하여 얻어진 자료를 분석하여 최적의 클러스터를 선정하는 방법은 2) 및 3)의 과정에서 많은 계산 시간이 소요되므로 비효율적인 방법이라 할 수 있다. 본 논문에서는, 이와 같은 단점을 보완하기 위하여 자료가 갖는 개략적 특성을 파악하고 이로부터 자료 속에 존재하는 클러스터의 근사적 개수 및 중심을 정한 후, 이 정보를 기존의 일반적인 퍼지 클러스터링 알고리즘에 입력하여 클러스터링을 수행하는 밀도함수를 이용한 계층적 구조의 근사적 클러스터링 알고리즘을 제안하고, 예제를 통하여 제안된 알고리즘의 타당성을 보인다. 모의 실험 결과로부터 제시된 계층적 구조의 근사적 클러스터링 알고리즘은 소속도 함수 (혹은 클러스터 중심)의 초기 값으로 임의의 값을 설정하는 기존의 퍼지 클러스터링 알고리즘에 비하여 계산 시간이 현저히 적으면서 성능 저하가 비교적 적은 특징을 가짐을 알 수 있다.

2. 밀도함수를 이용한 계층적 구조의 근사적 퍼지 클러스터링

이 절에서는 본 논문에서 제안하는 밀도함수를 이용한 계층적 구조의 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 여러 개의 부분 공간으로 분할하여 부분 공간 속에 속하는 자료의 개수를 자료의 밀도(density)로 정의한다면 클러스터 중심주변의 밀도가 다른 부분 공간의 밀도보다 높음을 알 수 있다. 즉, 밀도가 높으면 높을수록 클러스터의 중심이 될 가능성이 높다고 할 수 있다. 자료가 갖는 이와 같은 분포 특성을 이용하여 근사적인 클러스터의 수 및 클러스터 중심을 결정하고 이를 일반적인 퍼지 클러스터링 알고리즘의 초기 자료로 사용함으로써, 기존의 퍼지 클러스터링 알고리즘이 갖는 단점을 보완하고자 하는 것이 본 논문에서 제시하는 계층적 구조를 갖는 근사적 퍼지 클러스터링 방법이다.

이러한 밀도 함수를 이용한 근사적 퍼지 클러스터의 중심을 구하는 과정에서 중요한 문제중의 하나는 주어진 자료 공간의 분할과 분할된 자료들로부터 밀도를 계산하는 것이다. 본 논문에서 제시하는 클러스터링 방법은 Yager등의 산 클러스터링 방법에서와 같이 주어진 자료 공간을 부분공간으로 분할하고 이를 바탕으로 미리 정해진 목적 함수 값을 계산하여 클러스터의 수 및 클러스터 중심을 구한다는 점에서 유사하다고 할 수 있다. 그러나, 산 클러스터링 방법은 자료 공간을 임의로 분할하고 분할하는 선들의 교점에 잠재적인 클러스터 중심이 위치한다고 가정하여 복잡하게 정의된 목적 함수를 계산하여 최대 값을 갖는 점을 클러스터의 중심으로 선정하고 이 점에 의한 영향을 배제한 다음 다시 목적 함수를 계산하여 클러스터 중심을 구하는 순차적 알고리즘이라 할 수 있다. 이에 반하여, 본 논문에서 제시하는 방법은 자료의 특성에 따라 분할된 영역 속에 포함되는 자료의 개수로부터 측정된 밀도를 이용하여 주어진 자료 속에 존재할 수 있는 전체 클러스터에 대한 근사적인 클러스터의 중심을 찾는다는 것이 Yager등의 방법과 본 논문에서 제시한 방법의 큰 차이점이라 할 수 있다.

이러한 근사적 클러스터 중심을 구하는 과정은 크

게 2개의 과정(Phases)으로 이루어지며, 그 첫 번째 과정, 즉, Phsae I은 다음의 6단계 과정을 통해 이루어진다. 여기서는, 고려 대상 자료의 차원을 알고리즘 설명의 편의상 2차원 공간에 한정하기로 한다 .

Phase I

Step 1: 주어진 모든 자료 $x_j = (x_{1j}, x_{2j})$, $j = 1, 2, \cdots, n$을 각 좌표 축 x_1 및 x_2 상에 투영시킨다.

Step 2: 각각의 축 상에 투영된 각 점에서의 이웃하는 점까지의 최소 거리를 구한 후, 그 중 최대값을 구한다.

$$\Delta d^i = \max_j [\min_k d^i_{kj}] \text{ (단, } d^i_{jk} = \|x_{ij} - x_{ik}\|, \; i = 1, 2, j, k = 1, \cdots, n, j \neq k) \quad (1)$$

Step 3: Step 2에서 얻어진 값 Δd^i에 적절한 상수 M(예: $M = 0.1, 0.2, \cdots, 0.9, 1.0, 2.0, \cdots, 10$)을 곱한 값 $M \times \Delta d^i = \Delta d^i_M$을 이용하여 각 좌표계를 다음과 같이 분할한다.

$$x_{i_{min}} + p \times \Delta d^i_M \leq x_i < x_{i_{min}} + (p+1) \times \Delta d^i_M, \quad (2)$$

여기서 p는 0보다 크거나 같은 정수, $x_{i_{min}}$는 축 x_1 및 x_2 상에 투영된 좌표값의 최소값을 나타낸다.

Step 4: 분할된 각 구간에서의 밀도를 구하고 이를 바탕으로 밀도 함수 곡선을 구한 후, 밀도 함수 곡선에 존재하는 산 모양의 정점 수를 구한다.

Step 5: M 값을 변화시켜 산 모양의 정점의 수가 1이 될 때까지 Step 3 및 4의 과정을 반복한다.

Step 6: M 값의 변화에 따른 산 모양의 정점 수를 나타내는 곡선을 바탕으로 M 값의 변화에도 불구하고 정점의 수가 변하지 않고 가장 오래동안 유지되는 정점의 수를 근사적 클러스터의 수로 결정하고 이때의 클러스터 중심을 구한다

Phase II

Step 7: Step 6에서 구해진 클러스터 중심을 일반적인 FCM의 클러스터 중심 초기치로 설정하여 FCM을 수행한다.

위에서 제시한 알고리즘의 핵심은 각 좌표축에 투영된 각 점에서 가장 가까운 점까지의 거리 중에서 가장 큰 값인 Δd^i에 적절한 상수 M($M = 0.1, 0.2, \cdots, 0.9, 1.0, 2.0, \cdots, 10$)을 곱한 값 $M \times \Delta d^i = \Delta d^i_M$을 이용하여 클러스터의 수를 결정하는 것이라 하겠다. 여기서, 상수 M의 값이 1보다 큰 자연수 및 1보다 작은 소수를 포함하는 것은 3절의 모의 실험에서 보이는 바와 같이 주어진 자료가 극심한 잡음으로 오염된 경우, 오염된 자료 또한 다른 어느 자료들과 관련을 갖는다는 기본 개념에 의하여 구한 Δd^i 값은 상당히 크게 되어 자료 전체가 오직 하나의 클러스터만을 형성하는 것처럼 판단될 가능성이 있으므로 이를 방지하기 위한 방법이라 할 수 있다. 이와 같은 과정을 거침으로써 자료 속에 존재할 수 있는 잡음의 영향을 줄이거나 혹은 잡음 그 자체를 또 다른 하나의 클러스터로 인식할 수 있도록 하는 방법을 제시할 수 있기도 하다. 이와 같은 개념은 Hirota 등[19]의 부가적 자료를 갖는 변형 FCM방법과 유사한 성질을 갖고 있으나, Hirota등의 방법은 자료를 분류하고자 하는 주체가 미리 자료의 성격을 파악하여 자료 속에 존재하는 잡음과 유사한 자료를 미리 선별하여 하나의 클러스터로 지정하는 것이므로 그 성격이 상당히 다르다 할 수 있다.

앞에서 제안된 클러스터링 알고리즘을 일반적으로 2개의 클러스터를 가지는 것으로 알려진 그림 1의 나비형 자료를 이용하여 설명하기로 한다. 먼저, 주어진 자료를 X축과 Y축에 대해 투영한 후 각 좌표계에 대한 이산 구간을 식 (1) 및 (2)를 이용하여 구한 후, 좌표 공간을 분할한다. 분할된 영역에서의 밀도를 계산하여 그림으로 나타내면 그림 2(a) 및 (b)와 같다.

그림 1의 나비형 자료는 균등한 분포를 가지므로 모든 d^i_{jk}은 1임을 알 수 있다. 따라서 각각의 이산구간을 1로 하여 알고리즘을 적용해 밀도 함수를 구하면 X축에서는 2개의 산 모양이 존재하게 되므로 2개

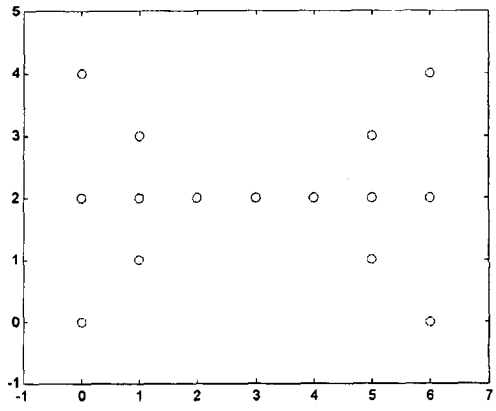
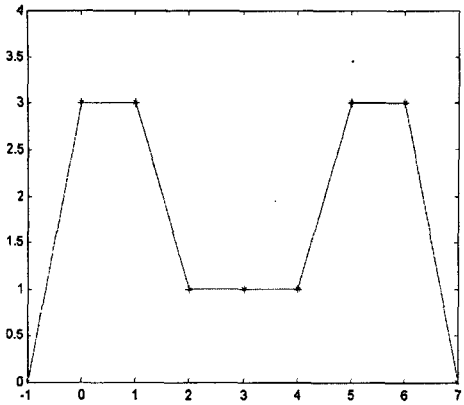
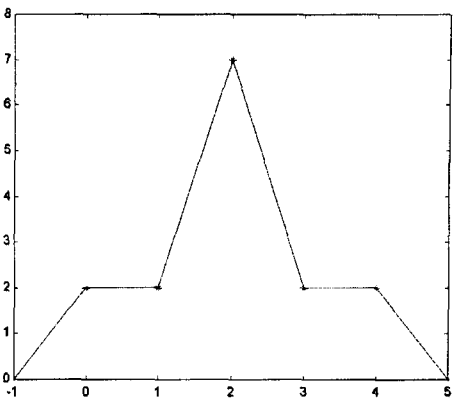


그림 1. 나비형 자료
Fig. 1. Butterfly data



(a) X축에 대한 각 구간에서의 밀도



(b) Y축에 대한 각 구간에서의 밀도

그림 2. 밀도 함수
Fig. 2. Density functions

의 클러스터가, *Y*축에서는 1개의 산 모양이 존재하므로 1개의 클러스터가 존재한다 할 수 있다. 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 갖는 것처럼 보인다. 이러한 문제를 해결하기 위해 이산 구간을 $M \times \Delta d^i$로 설정하여 다음의 모의실험을 통해 적절한 이산 구간을 구하는 방법을 보이고자 한다.

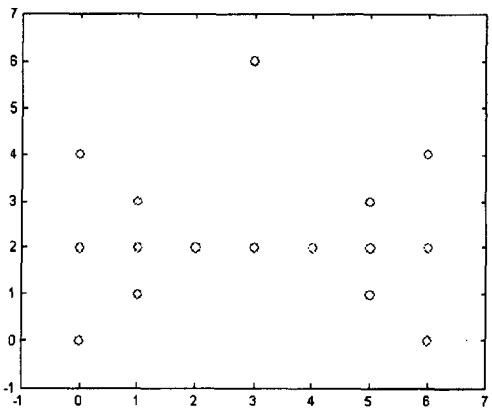
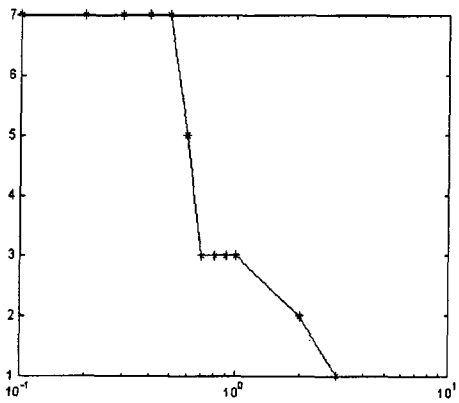
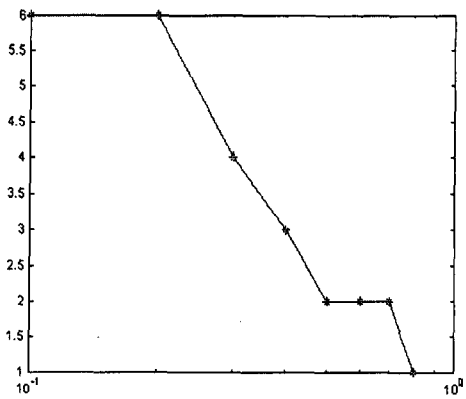


그림 3. 잡음이 섞인 나비형 자료
Fig. 3. Butterfly data with a noisy point



(a) X좌표



(b) Y좌표

그림 4. *M*값에 따른 클러스터 개수
Fig. 4. Number of clusters vs. M

3. 모의실험 결과 및 검토

이 절에서는 앞에서 제시한 알고리즘의 타당성을 보이기 위하여 그림 3에 나타난 잡음이 섞인 나비형 자료와 그림 6에 나타낸 불규칙적인 자료, 즉, 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 자료에 대한 모의 실험 및 결과를 나타낸다. 그림 3의 자료는 그림 1의 자료에 (3, 6)의 좌표를 갖는 점을 더한 자료로 클러스터링 알고리즘의 평가에 있어서 널리 사용되는 자료이다[19].

그림 4는 그림 3의 자료에 대하여 2절에서 제시한 알고리즘을 적용하여 얻어진 M값(횡축)의 변화에 따른 클러스터의 개수(종축)를 나타낸다. 그림 4의 (a)에서 알 수 있듯이 X좌표 자료를 7개의 클러스터로 분할하는 것은 X좌표 상의 하나 하나의 점들을 클러스터로 분할하는 것이므로, 이러한 경우를 배제하면 결국 3개의 클러스터가 존재함을 알 수 있다. 같은 방법으로 Y좌표 자료에 대하여 2절의 알고리즘을 적용하면 그림 4의 (b)에서와 같이 2개의 클러스터가 존재함을 알 수 있다.

위에서 얻어진 클러스터의 개수를 바탕으로 이에 해당하는 M값을 설정(X좌표 자료: M = 0.7, Y좌표 자료: M=0.5)하여 각각의 좌표공간을 분할하여 각각의 구간에서의 밀도를 구한 후 이를 그림을 나타내면 그림 5와 같다.

그림 5로부터 각각의 좌표에 대한 클러스터 중심 좌표를 구하여 2절에서 설명한 Phase II(일반적인 퍼지 클러스터링 알고리즘)를 수행한다. 그림 5의 모의 실험 결과로부터 본 논문에서 제시한 알고리즘으로 퍼지 클러스터링을 수행하는 경우, 잡음과 같은 성격을 지닌 자료를 본래의 자료로부터 분리할 수 있으므로, 임의의 자료에 대하여 클러스터링을 할 때 임의로 클러스터 수 및 중심을 설정하여 클러스터링을 수행하는 Hirota등이 제시한 알고리즘[19]보다 우수한 성능을 가짐을 알 수 있다.

다음으로는 그림 6에서 나타낸 5개의 클러스터 중심을 가지는 500개의 랜덤 자료, 즉, 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 자료에 대한 모의 실험 및 결과를 나타낸다.

그림 7은 각각의 좌표 공간에 대한 M값의 변화에 따른 밀도의 변화를 나타낸다. 그림 7에서 알 수 있듯이 M값이 작을수록 밀도 함수 곡선에는 보다 많은 산 모양이 나타나며 M 값이 크면 클수록 산 모양의

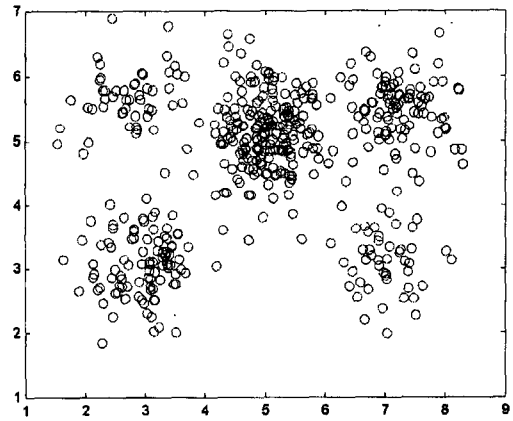
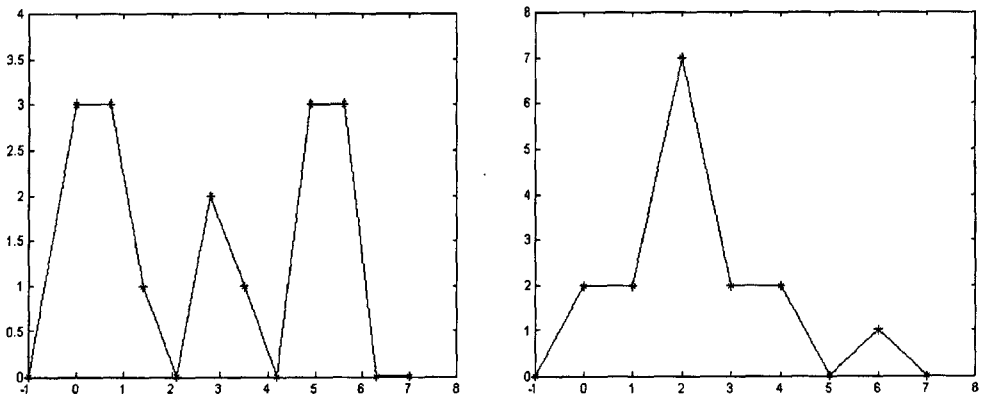


그림 6. 5개의 클러스터 중심을 가지는 500개의 랜덤 자료
Fig. 6. Five Gaussian clusters with 500 points



(a) X좌표 (b) Y좌표

그림 5. 밀도 함수
Fig. 5. Density functions

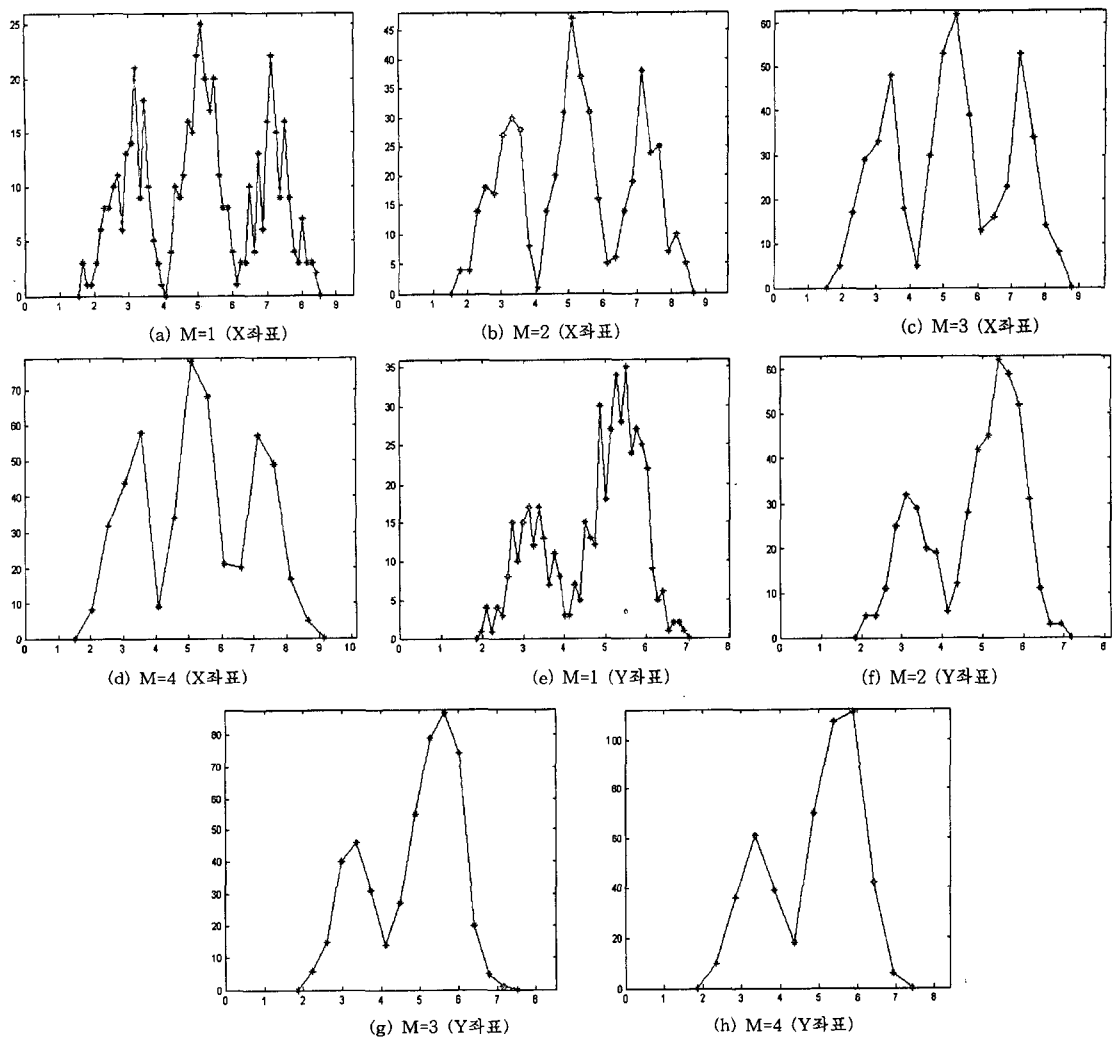


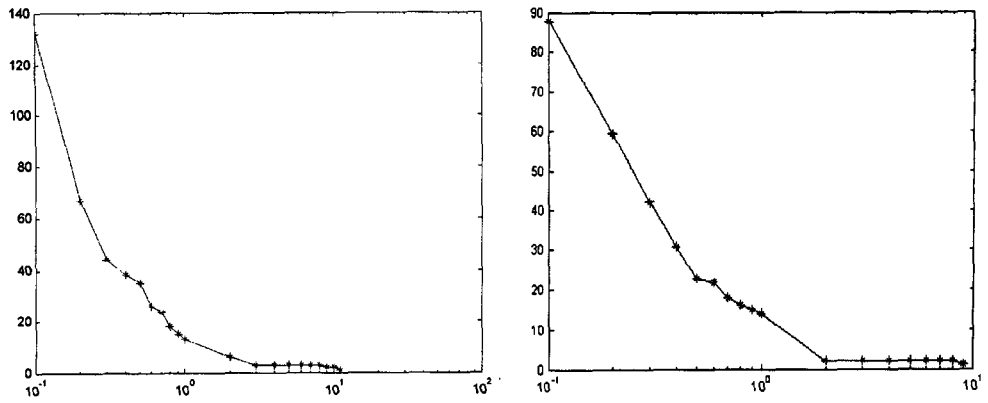
그림 7. 밀도 함수 곡선
Fig. 7. Density functions

수는 줄고 M값이 어느 값 이상이 되면 오직 1개의 산 모양만 나타나게 된다. 이것은 우리가 선험적으로 알 수 있는 바와 같이 상당히 큰 M값은 각각의 좌표 공간을 오직 하나의 부분 공간으로 분할하는 것과 같게 되어 오직 하나의 클러스터만이 존재하게 된다.

그림 8은 M값의 변화에 따른 클러스터의 개수를 나타내는 곡선으로, M의 값은 0.1에서 근사적인 클러스터 중심의 개수가 1이 될 때까지이다. 이 곡선을 보면 X좌표 자료의 경우 M이 3 그리고 Y좌표 자료의 경우 M이 2일 때부터 클러스터의 개수가 일정하게 유지된 후 1로 수렴함을 알 수 있다. 이 곡선으로부터 그림 6의 자료는 X좌표 자료의 경우 3개의 클러스터를 그리고 Y좌표 자료의 경우 2개의 클러스터를 가짐을 알 수 있다. M을 크게 하면, 즉 이산 구간이 커지면 보다 확실한 산 모양을 얻을 수 있지만 클러스터 중심 좌표의 정확성이 낮아지게 되므로, 본 논문에서는 근사적 클러스터 중심의 정확성을 높이기 위해 그림 7의 (c)와 (f)를 이용하여 근사적인 클러스터의 중심을 얻었다.

그림 9는 그림 6의 자료에 대해 본 논문에서 제시한 알고리즘과 일반적인 FCM 알고리즘을 적용하여 얻어진 클러스터의 중심 위치를 나타낸다.

그림 9와 표 1을 보면 그림 7(c)와 (f)에서 얻은 밀도를 이용하여 구한 근사적 퍼지 클러스터 중심은 그림 6에 나타낸 자료의 특성과 거의 일치함을 알 수 있다. 표 2는 일반적인 FCM과 본 논문에서 제안된



(a) X좌표 (b) Y좌표

그림 8. M값에 따른 클러스터 개수
Fig. 8. Number of clusters vs. M

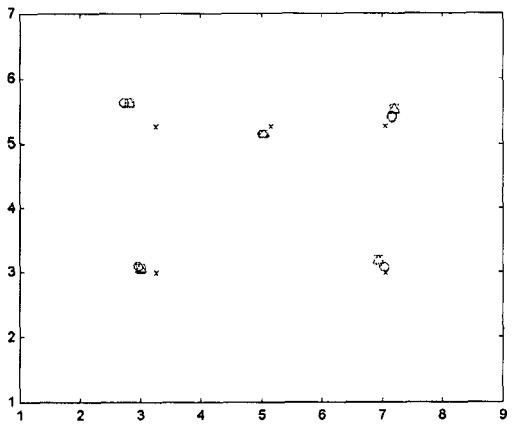


그림 9. 모의 실험 결과
Fig. 9. Simulation results

표 1. 본 논문에서 제시된 알고리즘과 일반적인 FCM의 클러스터 중심
Table 1. Cluster centers of the proposed and the regular FCM

클러스터 중심	일반적인 FCM 결과	Phase I	Phase II
(2.72, 5.62)	(2.8150, 5.6287)	(3.24625, 5.2691)	(2.8150, 5.6287)
(2.95, 3.09)	(2.9973, 3.0620)	(3.24625, 2.9903)	(2.9973, 3.0620)
(5.02, 5.13)	(5.0426, 5.1365)	(5.15275, 5.2691)	(5.0426, 5.1365)
(7.04, 3.07)	(6.9568, 3.1772)	(7.05925, 2.9903)	(6.9568, 3.1771)
(7.18, 5.41)	(7.2028, 5.5155)	(7.05925, 5.2691)	(7.2028, 5.5154)

알고리즘(Phase II)의 계산 시간을 비교한 것으로 본 논문에서 제시된 알고리즘을 사용하는 경우 계산 시간이 현저히 줄어듬을 알 수 있다. 단, Phase I은 프로그램의 알고리즘에 의해 계산 시간이 차이가 발생하므로 표 2에서는 고려하지 않았다.

표 2. 본 논문에서 제시된 알고리즘과 일반적인 FCM의 계산시간
Table 2. Computation time of the proposed and the regular FCM

	일반적인 FCM	Phase II
50회 평균	1.7037 sec	0.7024 sec
100회 평균	1.6765 sec	0.7036 sec
200회 평균	1.6450 sec	0.7018 sec

위의 3가지 수치 예제에 대한 모의 실험을 통해 보인 바와 같이, 본 논문에서 제시한 클러스터링 알고리즘, 즉, 자료의 성격을 바탕으로 클러스터의 개수 및 근사적 클러스터 중심을 찾아 이를 바탕으로 퍼지 클러스터링을 수행하는 계층적 구조를 갖는 클러스터링 알고리즘의 경우, 기존의 일반적인 클러스터링 알고리즘과 비교할 때 성능은 크게 변하지 않으면서도 계산 시간을 현저히 줄일 수 있으며 또한 동시에 클러스터의 개수에 관한 정보를 제공해 줌으로써 기존의 클러스터링 알고리즘이 갖는 여러 가지 단점을 보완하였음을 알 수 있다.

4. 결 론

본 논문에서는 다음의 3단계, 1) 자료가 갖는 경향 평가, 2) 클러스터 분석, 3) 클러스터의 타당성 조사로 이루어지는 자료 분석법에 있어서의 문제점, 즉, 단계 2) 및 3)의 반복 수행으로 인하여 많은 계산 시간이 소요되는 비효율성을 개선하기 위한 밀도함수를 이용한 계층적 구조의 근사적 퍼지 클러스터링 알고리즘을 제시하였다. 제시된 알고리즘은 자료가 갖는

개략적 특성을 파악하고 이로부터 자료 속에 존재하는 클러스터의 근사적 개수 및 중심을 정한 후, 이 정보를 기존의 일반적인 퍼지 클러스터링 알고리즘에 입력하여 클러스터링을 수행하는 계층적 구조를 갖는 근사적 클러스터링 알고리즘이다. 또한, 제안된 알고리즘의 타당성 및 유효성을 보이기 위하여 예제를 통한 모의 실험을 수행하였다. 모의 실험 결과로부터 제시된 계층적 구조의 근사적 클러스터링 알고리즘은 소속도 함수 (혹은 클러스터 중심)의 초기 값으로 임의의 값을 설정하는 기존의 퍼지 클러스터링 알고리즘에 비하여 계산 시간이 현저히 적으면서 성능 저하가 비교적 적은 특징을 가짐을 알 수 있었다.

향후 연구 과제로는 본 논문에서 제시된 클러스터 수 결정 알고리즘과 기존의 클러스터 타당성 평가 기준(Cluster validity index) 간의 연관성에 관한 이론적 고찰이 이루어져야 할 것이며, 또한 제시된 알고리즘의 실제 자료에 대한 적용을 통하여 제안된 알고리즘의 타당성 검증이라 할 수 있다. 위의 관계와 타당성이 검증된다면 본 논문에서 제시한 알고리즘은 보다 다양한 분야에 사용될 것으로 기대된다.

참고문헌

[1] G. W. Milligan, Clustering validation, *in Clustering and Classification*, P. Arabie, L. J. Hubert and G. D. Soete, Ed. World Scientific, Singapore, 1996.

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

[3] R. Krishnappuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, Vol. 1, No. 2, pp. 98-110, 1993.

[4] N. R. Pal, K. Pal and J. C. Bezdek, "A mixed c-means clustering model," in *Proc. FUZZ-IEEE97*, pp. 11-21, 1997.

[5] D. Titterington, A. Smith, and U. Markov, *Statistical Analysis of Finite Mixture Distributions*, New York, Wiley, 1985.

[6] R. R. Yager and D. P. Filev, "Approximate clustering via the mountain method," *IEEE Trans. Syst., Man, and Cybertn.*, Vol. 24, No. 8, pp. 1279-1284, 1994.

[7] R. L. Cannon, J. Dave and J. C. Bezdek, "Efficient implementation of the fuzzy c-means clustering algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 8, pp. 248-255, 1986.

[8] D. A. Linkens and M. -Y. Chen, "Hierarchical Fuzzy Clustering Based on Self-organizing Network," in *Proc. Fuzzy-IEEE'98*, pp. 1406-1410, 1998.

[9] T. W. Cheng, D. B. Goldg of and L. O. Hall, "Fast fuzzy clustering," *Fuzzy Sets and Syst.*, Vol. 93, pp. 49-56, 1998.

[10] H. Kalviainen, E. Oja and L. Xu, "Randomized Hough transform applied to translational and rotational motion analysis," in *Proc. 11th IAPR Internat. Conf. On Pattern Recognition*, Vol. I, Conference A: Computer vision and applications, IEEE Computer Soc. Press, Los Alamitos, pp. 672-675, 1992.

[11] J. C. Dunn, "Indices of partition fuzziness and the detection of clusters in large data sets," in *Fuzzy Automata and Decision Processes*, M. M. Gupta, Ed. Elsevier, New York, 1976.

[12] J. C. Bezdek and N. K. Pal, "Some new indexes of cluster validity," *IEEE Trans. Systems, Man, and Cyber-Part B*, Vol. 28, No. 3, pp. 301-315, 1998.

[13] X. L. Xie and G. A. Beni, "Validity measure for fuzzy clustering," *IEEE Trans. Pattern and Machine Intell.*, Vol. 3, No. 8, pp. 841-846, 1991.

[14] N. K. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Trans. Fuzzy Syst.*, Vol. 3, No. 3, pp. 370-379, 1995.

[15] S. H. Kwon, "Cluster validity index for fuzzy clustering," *Electronics Letters*, Vol. 34, No. 22, pp. 2176-2177, 1998.

[16] M. Delgado, A. F. Gomez-Skarmeta, and F. Martin, "A fuzzy clustering-based rapid prototyping for fuzzy rule-based modeling," *IEEE Trans., Fuzzy Syst.*, Vol. 5, pp. 223-233, 1997.

[17] M. Sugeno and T. Yasukawa, "A fuzzy logic based approach to qualitative modeling," *IEEE Trans, Fuzzy Syst.*, Vol. 1, No. 1, pp. 7-31, 1993.

[18] J. Yen, and L. Wang, "Application of statistical information criteria for optimal fuzzy model construction," *IEEE Trans. Fuzzy Syst.* Vol. 6, pp. 362-372, 1998.

[19] K. Hirota and K. Iwata, "Application of modified FCM with additional data to area division of images," Information Sciences, Vol. 45, pp. 213-230, 1988.

권 순 학 (Soon H. Kwon)

1983년 : 서울대학교 제어계측공학과 (공학사)

1985년 : 서울대학교 대학원 제어계측공학과(공학석사)

1995년 : 동경공업대학 시스템과학 (공학박사)

1996년~현재 : 영남대학교 전자정보공학부 부교수

관심분야 : 뉴시, 지능시스템 및 제어

손 세 호 (Seo H. Son)

2000년 : 영남대학교 전기전자공학부 (공학사)

2000년~현재 : 영남대학교 대학원 전기공학과 석사과정 재학중

관심분야 : 지능시스템 및 제어