퍼지 및 지능시스템학회 논문지 2001, Vol. 1, No. 1, pp. 1-8

개선된 산 클러스터링 방법

Advanced Mountain Clustering Method

이중우 · 손세호 · 권순학
Jung W. Lee, Seo H. Son, and Soon H. Kwon

영남대학교 전자정보공학부

요 약

본 논문에서는 정규화된 데이터 공간과 가우스함수에 의한 산 함수 형성 그리고 형성된 산의 기울기를 이용한 산봉우리 붕괴를 특징으로 하는 개선된 산 클러스터링 방법을 제안한다. 이 개선된 방법은 기존의 Yager등에 의하여 제안된 방법이 조정해야 하는 매개변수가 3 개이고 발견된 클러스터 중심 주위에 원치 않는 다른 중심이 발생할 수 있는데 반하여 단지 하나의 매개변수 ω의 조정으로 더욱 타당한 중심을 찾아내는 점에서 유용하다 할 수 있다. 또한 매개변수 ω에 대한 적절한 선정 방법을 제시하고, 수치 자료에 대한 컴퓨터 모의실험을 통하여 개선된 산 클러스터링 방법의 유용성을 입증한다.

ABSTRACT

We introduce an advanced mountain clustering method which uses a normalized data space, a gaussian type mountain function and a destruction method based on a mountain slope. This advanced method is more useful than Yager's mountain method because it needs just one parameter to tune instead of three and finds out more resonable cluster centers having no another centers neighboring to the first center. In addition, we present an adequate selection method for the only parameter ω. Finally, computer simulation results on numerical examples are presented to show the validity of the advanced mountain method.

Key Words : Clustering, Mountain Method, Advanced Mountain Method, Parameter Identification

1. 서 론

최근, 제어, 의사 결정, 패턴 인식, 지식 및 정보의 표현, 가공 및 전달 등등의 다양한 분야에서 비선형적이며 불확실성을 내포하면서도 매우 복잡한 구조를 갖는 시스템에 대한 관심이 점점 높아지고 있다. 또한, 이들에 대한 모델 식별 기법에 있어서도 기존의 수학적 모델을 사용한 기법으로부터 인간의 논리적 경험적 지식을 이용할 수 있는 퍼지 이론에 바탕을 둔 언어적 표현에 의한 모델링 방법이 많은 관심을 끌고 있다[1]. 이러한 퍼지 시스템은 퍼지 분할된 입력영역을 출력영역으로 대응시키기 위하여 If-then 규칙을 사용하며, 이 규칙은 전문가의 지식 혹은 시스템으로부터 직접 측정된 데이터로부터 얻어진다. 시스템을 식별하기 위한 규칙을 데이터로부터 얻을 때 많이 사용되는 방법으로는 유사도 척도를 이용한 방법[2], 통계적 지표를 이용하는 방법[10], 퍼지 클러스터링 및 학습을 통하는 방법 등등이 널리 이용되고 있다.

이와 같은 퍼지 시스템의 모델링 과정에서 일반적으로 사용되는 자료의 분석 과정은 1) 자료가 갖는 경향 평가, 2) 클러스터 분석, 3) 클러스터의 타당성 조사로 요약된다. 다시 말하면, 주어진 자료가 갖는 특성을 바탕으로 자료 속에 담긴 정보를 추출하여 이를 자료 분석에 이용하게 됨을 알 수 있다. 이러한 측면에서 퍼지 클러스터링은 주어진 데이터가 표현하는 시스템의 고유한 구조를 파악하기 위하여 퍼지이론을 도입하여 자료 집합을 데이터간의 거리등과 같은 유사한 특성을 가지는 그룹인 클러스터로 분할하는 기법으로 볼 수 있다.

일반적으로 많이 사용되는 퍼지 클러스터링 방법에는 Bezdek의 FCM(Fuzzy C-Means)[3]과 그의 단점을 보완하기 위한 개선된 방법들[4-5]이 있으나 이러한 방법들은 초기치의 설정이 어렵고 적절하지 못한 초기치의 설정은 클러스터링 과정과 클러스터의 타당성에 결정적인 영향을 미친다. 또한 선험적(heuristic) 접근법으로는 Yager와 Filev의 Mountain Method(MM)[6]와 이를 보완한 Subtractive 클러스터링[7]이 있으나 이 두 방법 모두 결정해야 할 매개변수(α, β, δ)가 많으며 적절치 못한 매개변수의 선정은 클러스터 중심 주위에서 여러 개의 중심이 생기는 등 좋지 않은 클러스터링 결과를 초래한다. 그리고 MM이 갖는 단점인 차원이 증가함에 따라

접수일자 : 2000년 11월 18일
완료일자 : 2001년 01월 15일
본 연구는 2000학년도 영남대학교 학술연구조성비 지원 받았습니다.

계산량이 지수함수적으로 증가하는 단점을 극복하기 위하여 격자선으로 분할된 입력 공간의 각각의 교점에서 클러스터 중심을 계산하지 않고 직접 각각의 자료로부터 클러스터 중심을 구하여 계산량을 줄인 Subtractive 클러스터링 알고리즘[7]이 제안되기도 하였으나 자료의 분포가 소(疎)한 영역에서의 클러스터링 결과가 좋지 않은 단점이 있다. 이 이외에도 밀도함수를 이용한 클러스터링 방법[8]이 제안되었으나 클러스터 재구성에 있어서 주어진 자료의 차원이 증가함에 따라 계산량이 지수함수적으로 증가하는 단점을 지니고 있다.

본 논문에서는 MM이 가지는 많은 매개변수(α, β, δ)를 선정해야 하는 문제점을 개선하기 위하여 단 하나의 매개변수(ω) 만으로도 만족스러운 결과를 얻을 수 있는 개선된 산 클러스터링 방법(Advanced Mountain Method: AMM)을 제안하고, 이 매개변수(ω)에 대한 선정기준이 될 수 있는 방법으로 자료의 성질(밀도[8], 분산 등)을 이용한 매개변수 선정 알고리즘을 제시하며, 이의 타당성을 모의 실험을 통하여 보이고자 한다.

2. Yager와 Filev의 산 클러스터링 방법 (Mountain Method: MM)

Yager등에 의하여 제시된 산 클러스터링 방법은 데이터 공간을 여러 개의 격자로 분할하고, 분할된 격자에서 각 데이터까지의 거리를 정의된 목적함수에 의하여 계산한 후, 이 값이 최대가 되는 격자의 좌표를 클러스터 중심으로 선택하는 방법이다. 이렇게 하나의 중심을 선택한 후 다음 중심을 찾기 위하여 그 봉우리를 붕괴하는데 이때 붕괴용 목적함수를 새로이 정의하게 된다. 이러한 과정은 무한히 반복 될 수 있으므로 최초로 선택된 클러스터 중심의 산 높이와 현재의 산 높이를 비교하여 종료여부를 결정하게 된다. 결국 산을 형성하기 위한 목적함수, 봉우리를 붕괴하기 위한 목적함수, 종료조건을 결정하기 위한 산 높이의 비율 각각에 대하여 매개변수를 필요로 하므로 클러스터링을 위하여 세 개의 매개변수(α, β, δ)의 설정을 필요로 한다. Yager등의 산 클러스터링 방법을 살펴보면 다음과 같다.

주어진 s차원 공간상의 n개의 데이터 $\{x_1, \cdots, x_n\}$를 포함하는 최소의 공간 I_j를 구성한후

$$I_j = [\min_k(x_{kj}), \max_k(x_{kj})]\text{단, j=(1,...,s), k=(1,...,n)} \quad (1)$$

(여기서, x_{kj}는 k번째 데이터 x_k의 j번째 좌표축으로의 투영이다.)

I_j를 r_j개의 구간으로 나누고 그 격자 $\{X_1^{(j)}, \cdots, X_{rj}^{(j)}\}$가 만나는 점을 노드 $(N_i = N_{(i1,\cdots,is)})$라 하고 노드와 데이터와의 거리 $d(x_k, N_i)$를 구한다.

$$d(x_k, N_i) = (|x_{k1} - X_1^{(1)}|_p + |x_{k2} - X_3^{(2)}|_p)^{1/p} \quad (2)$$

식(2)에서 정의된 거리를 이용하여 노드 N_i에서의 산 높이 $M(N_i)$를 구한다.

$$M(N_i) = \Sigma_{k=1}^{n} e^{-\alpha \cdot d(x_k, N_i)} \quad (3)$$

이렇게 만들어진 산(Mountain)에서 산 높이가 최대가 되는 노드의 좌표를 중심으로 취한다.

$$M_1^* = Max_i[M(N_i)] \quad (4)$$

다시 그 노드의 산봉우리를 무너뜨리는 식(5) 과정을 반복하면서 중심을 찾아 나간다.

$$M^k(N_i) = \max[M^{k-1}(N_i) - M_{k-1}^* \Sigma_{k=1}^{n} e^{-\beta \cdot d(N_{k-1}^*, N_i)}, 0] \quad (5)$$

이러한 과정을 최대 산 높이에 비하여 충분히 산 높이가 작아질 때까지 반복한다.

$$\frac{M_1^*}{M_{k-1}^*} < \delta \quad (6)$$

결국, 산을 만들기 위해 α, 봉우리를 붕괴하기 위해 β, 종료 조건으로 δ의 3개의 매개변수값의 설정이 필요하다. Yager등이 제안한 MM은 많은 매개변수 설정문제와 매개변수의 설정에 따라 클러스터링의 결과가 크게 변하는 단점을 지니고 있음을 알 수 있다.

3. 개선된 산 클러스터링 방법 (Advanced Mountain Method: AMM)

Yager등의 MM은 많은 장점에도 불구하고 결정해야 하는 매개변수가 많아 실제 사용에는 많은 제약을 가지게 되고, 잘못된 매개변수 설정은 데이터의 밀도가 상대적으로 높은 클러스터의 중심 주위에 원치 않는 다른 클러스터 중심을 발생시킬 가능성을 내포하고 있다. 이러한 Yager등의 MM이 갖는 단점을 개선하기 위하여 본 논문에서는 다음과 같은 3가지 측면에서 개선된 산 클러스터링 방법(Advanced Mountain Method: AMM)을 제안한다.

그 첫 번째로는, 제안된 방법에서는 정규화된 데이터 공간을 격자로 분할하여 사용함으로써 동일한 구조를 가지고 있는 데이터에 대하여 동일한 매개변수 값으로 적절한 클러스터링 결과를 얻을 수 있어 매우 유용하다 할 수 있다. 두 번째로는, MM에서는 산봉우리 형성을 위한 목적함수로 지수함수를 사용하는데, 이는 충분히 가까운 거리에 위치한 데이터가 봉우리 높이는 크게 차이가 생겨서 직관적으로 타당하지 않을 뿐만 아니라 산 모양이 불연속적인 형태가 되어 취급에 많은 제한이 있으나, AMM에서는 가우시안(Gaussian) 형태의 함수를 사용하여 이러한 문제를 제거한다. 그리고 마지막으로 MM에 있어서 산봉우리를 붕괴하며 클러스터 중심을 찾는 과정에서 새로운 목적함수의 설정이 필요하고 상대적 데이터의 밀도차에 의한 2차적인 원치 않는 중심이 생기는 문제를 해결하기 위하여 산의 기울기를 이용하는 새로운 방법을 제안한다.

이러한 방법을 사용함으로써 Yager등의 MM에서 필요한 3개의 매개변수(즉, 산을 만들기 위한 α, 봉우리를

붕괴하기 위한 β, 종료 조건으로의 δ) 대신에, 산을 만들기 위한 매개변수 ω만을 설정하는 것이 요구되는 매우 단순하면서도 유용한 개선된 산 클러스터링 방법을 제시한다.

개선된 산 클러스터링 알고리즘을 살펴보면, 주어진 s 차원 공간 $\boldsymbol{R}^s$상의 n개의 데이터를 포함하는 최소공간 I_j를 식(1)과 동일하게 구성한 후 r_j개의 구간으로 나눈 후 노드 N_i와 데이터 x_k의 거리를 정규화(Normalized)된 Euclidean Norm으로 구한다.

$$D_j = \max_k(x_{kj}) - \min_k(x_{kj})\text{단, j=(1,...,s), k=(1,...,n)} \quad (7)$$

$$d_{\max} = (\Sigma_{j=1}^{s} D_j^2)^{\frac{1}{2}} \quad (8)$$

$$d_n(x_k, N_i) = \frac{||x_k - N_i||_2}{d_{\max}} \quad (9)$$

(여기서, x_{kj}는 k번째 데이터 x_k의 j번째 좌표축으로의 투영이다.)

식(3)의 지수함수는 근접한 데이터에 대하여 산봉우리의 높이가 크게 차이가 나서 근접한 데이터를 동일한 클러스터로 취급하는 클러스터링의 본래의 취지에 벗어나므로 근접한 데이터에 대하여 비슷한 산봉우리를 형성하기 위하여 가우시안 함수를 적용하여 정규화된 거리로 가우시안 함수에 의하여 노드의 산 높이 $M(N_i)$를 구한다.

$$M(N_i) = \Sigma_{k=1}^{n} e^{-d_n(x_k, N_i)^2/2w^2} \quad (10)$$

산 높이가 최대가 되는 점의 좌표를 식(4) $M_1^* = Max_i[M(N_i)]$에 의하여 첫 번째 클러스터 중심으로 취한 후 산 높이가 다시 높아지거나 바닥($M(N_i)=0$)이 되는 노드를 만날 때까지 반복하면서 산을 내려온다. 이와 같이 산의 기울기에 의해 산을 붕괴함으로써 클러스터를 형성시킨다.

위의 AMM알고리즘을 요약하면 다음과 같다.

Step 1 : 식(1)을 이용하여 구간 I_j를 계산한다.

Step 2 : 구간 I_j를 나누어 격자를 형성한다.

Step 3 : 식(10)을 이용하여 산 높이를 계산한다.

Step 4 : 산 높이가 최대인 점에서 클러스터 중심을 구한 후, 산 높이가 다시 높아지거나 바닥($M(N_i)=0$)이 되는 노드를 만날 때까지 산을 내려오면서 산봉우리를 붕괴한다.

Step 4-1: 산 높이가 최대 $Max(M(N_i))$ 인 N_i ($=N_{(i1,\cdots,is)}$)를 찾는다.

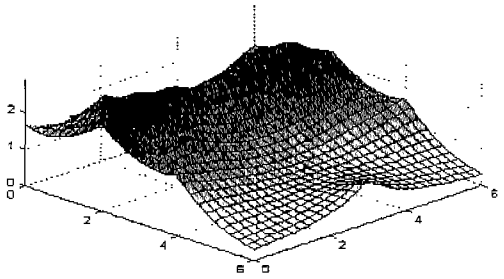
Step 4-2: 인접한 노드 N_i^* ($=N_{(i1+1,\cdots,is)}$, $N_{(i1-1,\cdots,is)}$, $N_{(i1,\cdots,is+1)}$, $N_{(i1,\cdots,is-1)}$) 각각에 대하여 Step 4-3, 4를 반복한다.

Step 4-3: $M(N_i^*) > M(N_i)$ 또는 $M(N_i^*)=0$이면 재귀호출(Recursive call)을 중단한다.

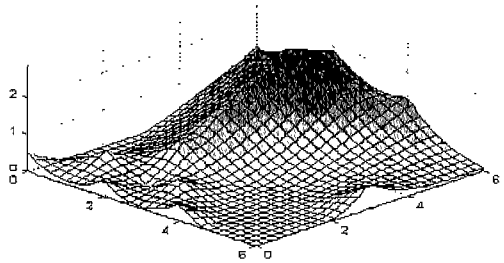
Step 4-4: $M(N_i^*) \leq M(N_i)$ 이면 $N_i = N_i^*$로 두고 Step 4-2를 재귀호출 한후 $M(N_i^*)=0$로 한다.

Step 5: 모든 노드의 산 높이가 영이 될 때까지 Step 4를 반복한다.

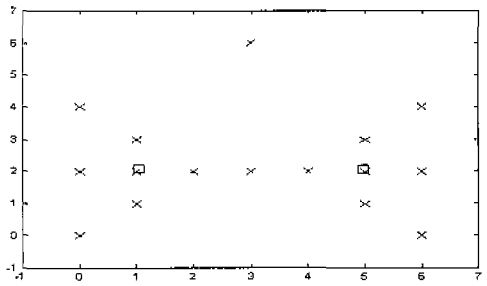
AMM은 위의 알고리즘을 통해 알 수 있듯이 산봉우리를 만들 때 식(10)의 ω 하나만을 조절하여 적절한 중심을 구할 수 있으므로 유용할 뿐만 아니라 정규화된 거리 $d_n(x_k, N_i)$를 사용하므로 $0 < \omega < 1$ 이 되고, 동일한 패턴의 확대 및 축소된 데이터에 대하여도 동일한 ω값을 가진다.



(a) MM에 의한 산모양(α =1)



(b) 봉우리 붕괴 후(β =0.4)

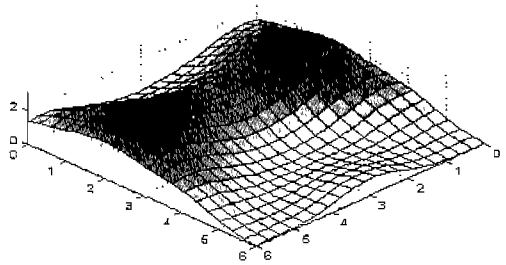


(c) 클러스터 중심(δ =0.2)

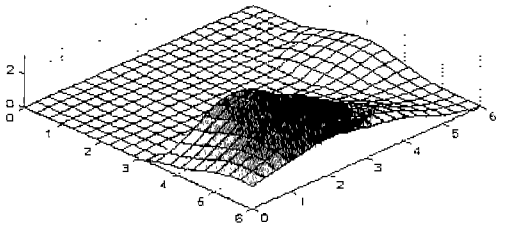
그림 1. 산 클러스터링 방법

Fig. 1. Mountain Method

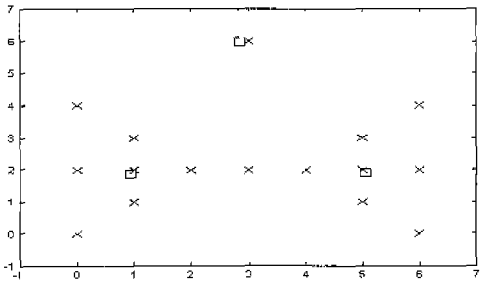
이러한 사실들은 그림 1 및 그림 2를 보면 확연히 알 수 있다. Yager와 Filev가 제시한 MM에 의한 산봉우리(그림 1.a)는 그 모양이 볼록(Convex)하지 않으며, 봉우리가 붕괴(Destruction)된 이후에도 그 주위에 다시 낮은 2차 봉우리가 나타남으로 인해 클러스터 중심 주위에 불필요한 또 다른 중심이 나타날 가능성을 포함하고 있다. 그리고 최적의 클러스터 중심을 찾기 위해 α, β, δ



(a) AMM에 의한 산모양(ω=0.1)



(b) 봉우리 붕괴 후(ω=0.1)



(c) 클러스터 중심

그림 2. 개선된 산 클러스터링 방법

Fig. 2 Advanced Mountain Method

세 개의 매개변수를 조정하여야 하므로 최적의 매개변수 선정을 위해 지나치게 많은 반복수행을 필요로 함을 알 수 있다. 그러나 본 논문에서 제시하는 AMM에 의한 산봉우리(그림 2.a)는 볼록한 봉우리를 형성할 가능성이 높으며, 1차로 발견된 클러스터 중심의 주위에 2차 봉우리를 형성하지 않으며 또한 조정하여야 할 매개변수가 오직 ω 한 개뿐이므로 조정이 간편하고 그 범위가 넓어(실제로 위의 데이터의 경우 $\omega = 0.08 \sim 0.14$에서 적절한 중심을 찾아냄) 보다 적은 반복 수행에 의한 클러스터링 알고리즘의 가능성을 보여주고 있다.

4. AMM의 매개변수 ω 설정법

앞의 2절과 3절에서 언급하였듯이 MM이 가지는 α, β, δ 매개변수 설정 문제를 AMM에서는 단 하나의 매개변수 ω를 선정하는 문제로 개선하였지만 여전히 적절한 매개변수 ω를 선정하여야 하는 문제점을 가지고 있다. 이 절에서는 자료들이 가지고 있는 특성(밀도, 분산 등)을 이용하여 적절한 ω를 선정하는 알고리즘을 제시하고자 한다.

주어진 자료들이 가지는 특성(자료의 밀도)을 바탕으로 자료 공간을 분할하여 얻어진 각각의 부분 공간에서의 분산을 구한다. 분할된 각각의 부분 공간상에는 단 하나의 클러스터가 존재하기 때문에 분할된 공간상에 존재하는 자료들의 분산은 자료의 흩어진 정도를 나타내므로 클러스터의 크기를 결정하는 중요한 요인이 된다. AMM에서는 산을 만들기 위해 가우시안 함수를 사용하므로 일반적으로 다음과 같은 성질을 만족하게 된다. 즉, 분산이 작으면 작을수록 클러스터의 크기는 작아지고 분산이 커지면 커질수록 클러스터의 크기는 커지게 되며, 분할된 공간의 분산 중 가장 작은 값을 사용함으로써 분류하고자 하는 최소의 클러스터의 크기를 결정할 수 있다. 이와 같은 특성을 바탕으로 구성한 매개변수 설정법은 다음과 같이 Phase I과 Phase II로 구성되어 지며 Phase I은 자료 공간 분할에 관련된 알고리즘을 그리고 Phase II는 분할된 각각의 공간에서의 자료들이 갖는 분산을 바탕으로 ω를 선정하는 알고리즘을 나타낸다.

Phase I [8] .

Step 1: 주어진 모든 자료 $x_j = (x_{1j}, x_{2j})$ j=1,2, ...,n 을 분산이 1이 되도록 정규화 시킨다.

$$x_{j,\ normalized} = x_j / N_{max} \ ,$$

$$N_{max} = \left(\sum_{k=0}^{s} n_k^2 \right)^{1/2},$$

$$n_k = \max_s (x_{sj}) - \min_s (x_{sj}) \quad s=1,2,\ j=1,2,...,n$$

Step 2: 정규화된 자료 $x_{j,normalized}$를 각 좌표 축 x_1 및 x_2 상에 투영시킨다.

Step 3: 각각의 축 상에 투영된 각 점에서의 이웃하는 점까지의 최소 거리를 구한 후, 그 중 최대값을 구한다.

$$\Delta d^s = \max_j \left[\min_k d^s_{jk} \right] \qquad (11)$$

(단, $d^s_{jk} = ||x_{sj} - x_{sk}|| s=1,2,\ j, k=1, ..., n, j \neq k$)

Step 4: Step 3에서 얻어진 값 Δd^s에 적절한 상수 M (예: M = 0.1, 0.2, ..., 0.9, 1.0, 2.0, ..., 10)을 곱한 값 $M \times \Delta d^s = \Delta d^s_M$을 이용하여 각 좌표계를 다음과 같이 분할한다.

$$x_{s_{min}} + p \times \Delta d^s_M \leq x_s < x_{s_{min}} + (p+1) \times \Delta d^s_M, \qquad (12)$$

(여기서 p는 0보다 크거나 같은 정수, $x_{s_{min}}$는 축 x_1 및 x_2상에 투영된 좌표값의 최소값을 나타낸다.

Step 5: 분할된 각 구간에서의 밀도를 구하고 이를 바탕으로 밀도 함수 곡선을 구한 후, 밀도 함수 곡선에 존재하는 산 모양의 정점 수를 구한다.

Step 6: M 값을 변화시켜 산 모양의 정점의 수가 1이 될 때까지 Step 4 및 5의 과정을 반복한다.

Step 7: M 값의 변화에 따른 산 모양의 정점 수를 나타내는 곡선을 바탕으로 M 값의 변화에도 불구하고 정점의 수가 변하지 않고 가장 오래 동안 유지되는 정점의 수를 근사적 클러스터의 수로 결정한다.

Phase II

Step 8: Step 7에서 결정된 클러스터를 바탕으로 공간을 분할한 후 본래의 차원으로 확장시킨다.

Step 9: Step 8에서 생성된 각 부분공간에서의 자료 $x_i \in x_j$, i=1,2, ..., n′ 에 대한 대표값 (μ_kk=1,2,...)과 분산(σ_k,k=1,2,...)을 구한다.

$$\mu_k = \frac{1}{n'} \sum_{i=1}^{n'} x_i ,$$

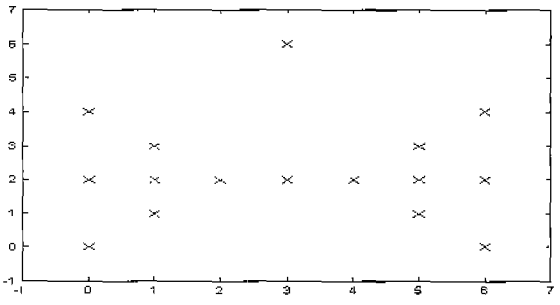
$$\sigma_k = \frac{1}{n'} \sum_{i=1}^{n'} \|x_i - \mu_k\|_2^2$$

(n′>1이며 공간 분할선에 존재하는 자료는 두 공간 모두에 소속시킨다.)

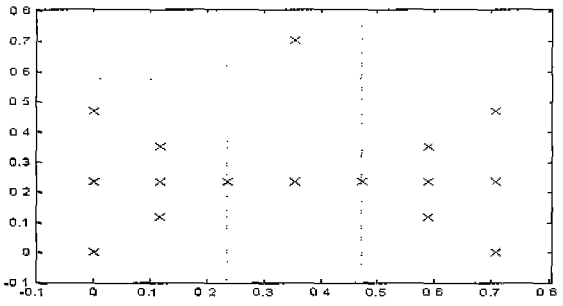
Step 10: ω를 선택하여 AMM을 수행한다.

$$\omega = \min_k (\sigma_k)^{1/2}, k = 1, 2, \ldots \qquad (13)$$

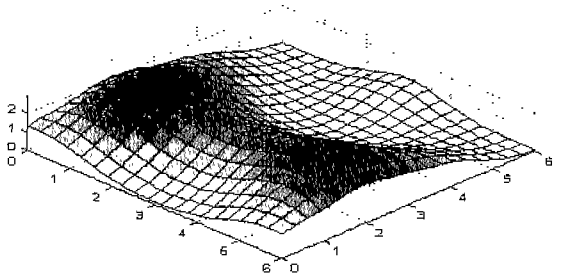
위에서 제시한 알고리즘을 2절과 3절에서 비교한 자료에 적용시켜 보면 그림 3과 같은 결과를 얻을 수 있다. 그림 3.(b)에서 보면 주어진 자료는 Phase I에 의해 6개의 공간으로 분할되며 Phase II에서 분할된 각 공간에서의 분산을 구하면 σ_1^2=0.0266, σ_2^2=0.0093, σ_3^2=0.0266 이다. 이 중 가장 작은 값으로 설정한 ω=0.0964를 사용하여 AMM을 수행한 결과는 그림 3.(c),(d)와 같다.



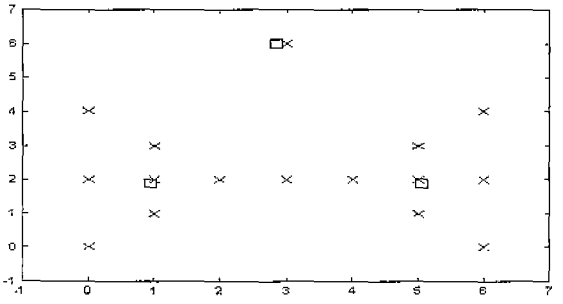
(a) 잡음이 섞인 나비형 자료



(b) 밀도에 의해 분할된 공간



(c) AMM에 의한 산모양 (ω=0.0964)



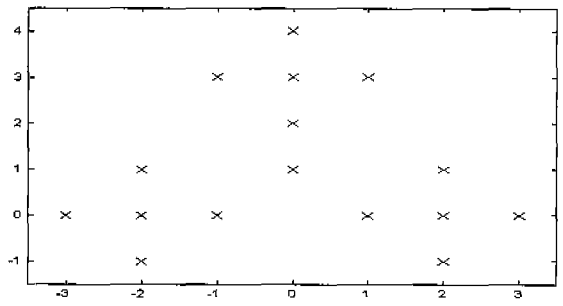
(d) AMM 수행된 결과

그림 3. AMM의 매개변수 설정

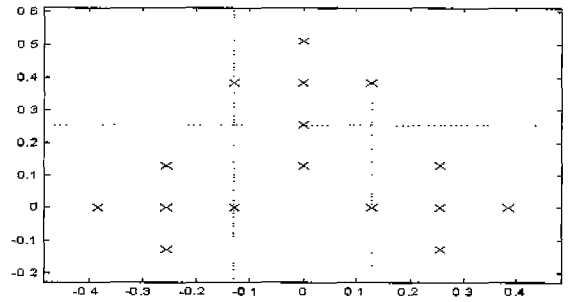
Fig. 3 Selection of AMM's parameter

5. 모의실험 결과 및 검토

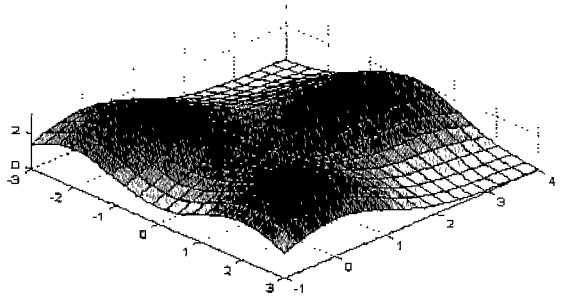
이 절에서는 본 논문에서 제시한 AMM 알고리즘과 매개변수 설정법에 대한 타당성을 보이기 위하여 그림 4 및 5와 같이 규칙적인 형태를 가지는 자료, 그림 6과 같이 임의의 다섯 점 (2.72,5.62), (2.95,3.09), (5.02,5.13), (7.04,3.07), (7.18,5.41) 주위에 분산 0.5로 분포한 500개의 랜덤자료[8], 마지막으로 그림 7과 같이 Yager등의 논문에서 사용된 자료[6]에 대하여 모의 실험을 수행하였다.



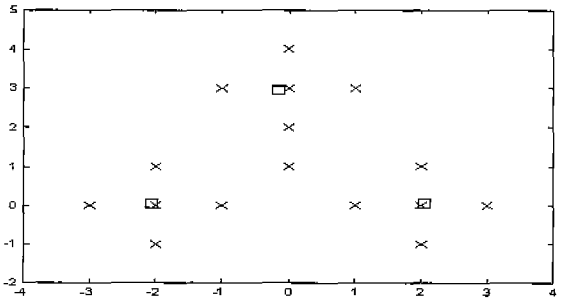
(a) 십자가형 자료



(b) 밀도에 의해 분할된 공간

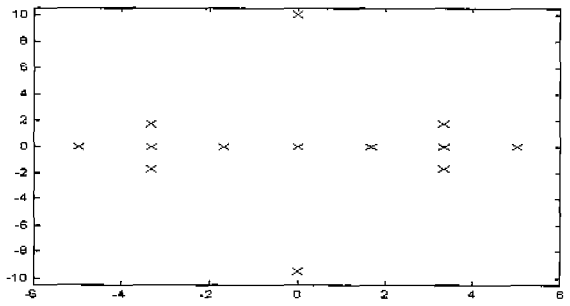


(c) AMM에 의한 산 모양 (ω=0.1145)

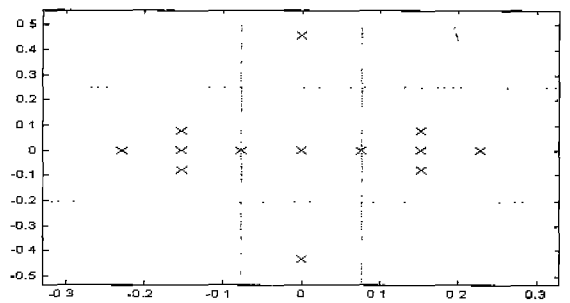


(d) 모의실험 결과

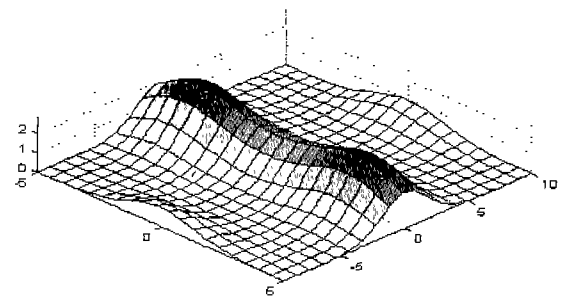
그림 4. 규칙적인 형태를 갖는 자료(Ⅰ)
Fig. 4 Simulation Data Ⅰ without noise



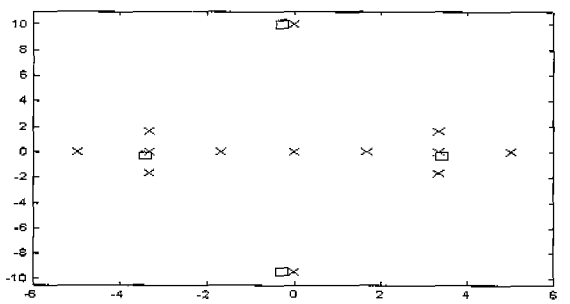
(a) 주어진 자료



(b) 밀도에 의해 분할된 공간



(c) AMM에 의한 산모양(ω=0.0622)

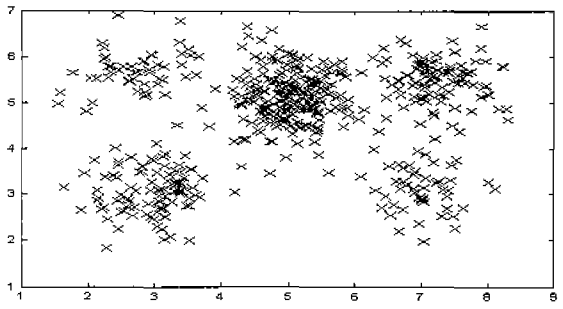


(d) 모의실험 결과

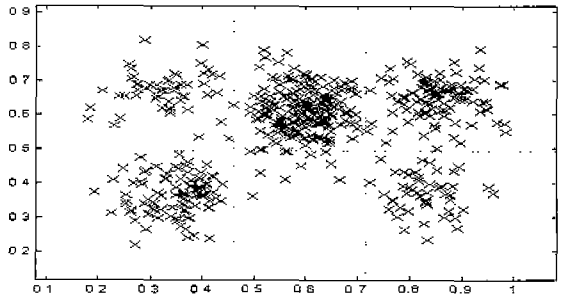
그림 5. 규칙적인 형태를 갖는 자료(Ⅱ)
Fig. 5 Simulation Data Ⅱ without noise

그림 4의 십자가형 자료는 Phase Ⅰ에 의해 그림 4.(b)처럼 6개의 부분공간으로 나누어진다. Phase Ⅱ에서 각각의 부분공간에 대한 분산을 구하면 σ_1^2=0.0131, σ_2^2=0.0195, σ_3^2=0.0131, σ_4^2=0.0131이다. 그림 4.(c),(d)는 Phase Ⅱ에서 설정한 ω=0.1145로 AMM을 사용한 결과이며 AMM이 적절한 클러스터 중심을 찾아내는 것을 알 수 있다.

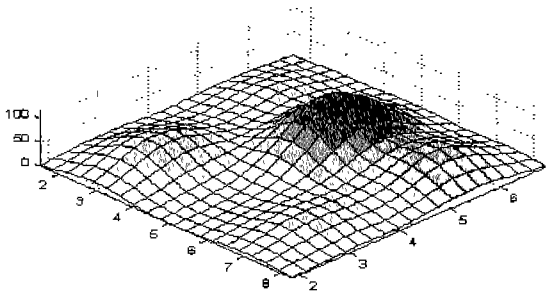
그림 5의 자료에 대해 Phase Ⅰ은 주어진 자료 공간을 9개의 부분공간(그림 5.(b))으로 분할하며 Phase Ⅱ에서 각각의 부분공간의 분산을 구하면 σ_1^2=0.0046, σ_2^2=0.0039, σ_3^2=0.0046이다. 그림 5.(c),(d)는 Phase Ⅱ에서 설정한 ω=0.0622로 AMM을 수행한 결과이다.



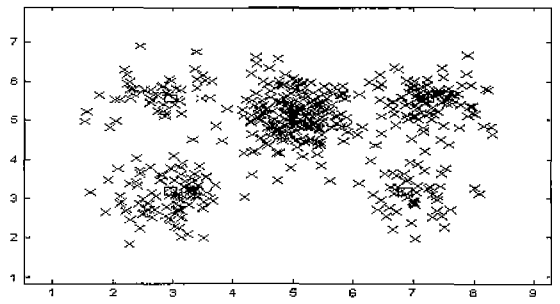
(a) 500개의 랜덤 자료



(b) 밀도에 의해 분할된 공간



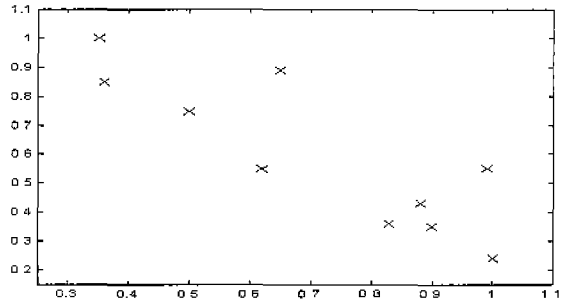
(c) AMM에 의한 산 모양(ω=0.0702)



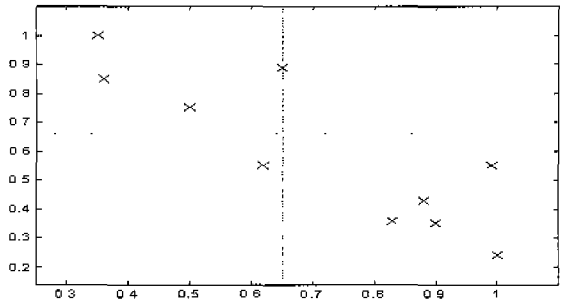
(d) 모의실험 결과

그림 6. 불규칙적인 형태를 갖는 자료(Ⅲ)
Fig. 6 Simulation Data Ⅲ with noise

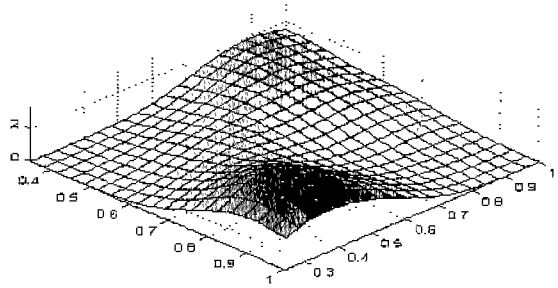
그림 6의 다섯 개의 점 (2.72, 5.62), (2.95, 3.09), (5.02, 5.13), (7.04, 3.07), (7.18,5.41) 주위에 분산 0.5로 분포한 500개의 랜덤자료는 Phase I에 의해 6개의 부분공간(그림 6.(b))으로 분할된다. Phase II에서 각각의 부분공간의 분산은 σ_1^2=0.0060, σ_2^2=0.0077, σ_3^2=0.0049, σ_4^2=0.0061, σ_5^2=0.0055, σ_6^2=0.0068이며 적절한 ω는 $0.0049^{1/2}$가 된다. 그림 6.(c),(d)는 Phase II를 통해 설정한 ω=0.0702를 사용한 AMM이 근사적인 클러스터 중심 (2.96,5.57), (2.96,3.18), (5.10,5.04), (6.88,3.18), (7.23,5.57)을 찾아내는 것을 보여주고 있다.



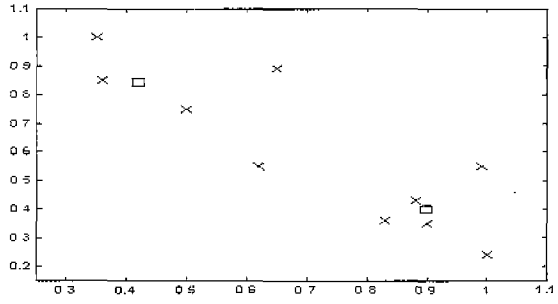
(a) Yager 등의 논문에서 사용된 자료



(b) 밀도에 의해 분할된 공간



(c) AMM에 의한 산모양(ω=0.1212)



(d) 모의실험 결과

그림 7. 불규칙적인 형태를 갖는 자료(Ⅳ)
Fig. 7 Simulation Data Ⅳ with noise

마지막으로 그림 7에서 보인 Yager등의 논문에서 사용된 자료는 Phase I에 의해 4개의 부분공간(그림 7.(b))으로 나누어진다. ω는 Phase II를 통해 각각의 부분공간의 분산 σ_1^2=0.0229, σ_2^2=0.0148 중 가장 작은 분산의 제곱근(0.1217)이 된다. 그림 7.(c),(d)를 보면 Yager등의

MM으로 찾은 클러스터 중심 (0.888,0.371), (0.372,0.843) 보다 본 논문에서 제시한 AMM으로 찾은 클러스터의 중심 위치 (0.897,0.400),(0.418,0.840)가 직관적으로 더욱 타당함을 알 수 있다.

6. 결 론

본 논문에서 제시한 개선된 산 클러스터링 방법(AMM)은 Yager등의 논문[6]의 결론부에서 제시한 매개변수 α, β, δ의 선택에 의한 클러스터링 결과의 민감성에 대한 문제를 매개변수를 ω 한 개로 줄임으로써 부분적으로 해결하였다. 즉, MM의 산봉우리 형성함수를 지수 함수형에서 매개변수로 ω만을 갖는 가우시안 함수형으로 선정하고 산봉우리 붕괴 알고리즘을 산봉우리의 경사를 바탕으로 구축하여, 먼저 생성된 클러스터 중심 주위에서 발생할 수 있는 기생 클러스터의 발생 가능성을 제한하였다. 그리고, 노드와 데이터의 거리를 정규화하여 매개변수 ω의 존재범위를 제한하였다. 또한 주어진 자료의 밀도에 의해 분할된 부분 공간들의 분산을 이용한 매개변수 ω의 설정법을 제시하여 제안된 산 클러스터링 방법(AMM)이 보다 우수한 성능을 가지면서도 효율적인 클러스터링 알고리즘임을 보였다.

향후 과제로는 모의 실험에서 보인 클러스터링 결과와 기존의 Cluster validity index[9]를 사용한 결과와의 비교 검토를 통하여 클러스터의 타당성 검증이 이루어져야 하며, 또한 부분 공간상의 분산과 매개변수 ω와의 연관성에 대한 이론적인 고찰이 필요하다고 보며, 제안된 알고리즘의 실제 자료에 대한 응용에 대한 연구가 필요하다고 본다. 마지막으로, MM이나 AMM이 갖는 단점인 차원이 증가함에 따라 계산량이 지수함수적으로 증가하는 단점을 극복하기 위하여 격자선으로 분할된 입력 공간의 각각의 교점에서 클러스터 중심을 계산하지 않고 각각의 자료로부터 직접 클러스터 중심을 구하는 방법으로의 적용 가능성에 관한 연구도 필요하다고 본다.

참 고 문 헌

[1] 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.

[2] M. Setnes, R. Babuska, U. Kaymak, and H. R. van N.Lemke, "Similarity measures in fuzzy rule base simplification," *IEEE Trans.*, SMC(B), Vol. 28, No. 3, pp. 376-386, 1998.

[3] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, NewYork, 1981.

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

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

[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] S. Chiu, "Fuzzy Model Identification Based on Cluster Estimation," *Journal of Intelligent & Fuzzy Systems*, Vol.2, No.3, Sept., 1994.

[8] 권 순학, 손 세호, "밀도함수를 이용한 근사적 퍼지 클러스터링," 한국 퍼지 및 지능 시스템학회 논문지, 제 10권, 제 4호, pp. 285--292, 2000.

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

[10] J. Yen, and L. Wang, "Application of statistical information criteria for optimal fuzzy model construction," *IEEE Trans., Fuzzy Syst.*, Vol. 6, No. 3, pp. 362-372, 1998.

저 자 소 개

이 중 우 (Jung W. Lee)
1991년 : 영남대학교 기계공학과 (공학사)
1991년~ 1996년 : 삼성중공업 연구원
2000년~ 현재 영남대학교 대학원 전기공학과 석사과정 재학중
관심분야 : 지능제어 등

손 세 호 (Seo H. Son)
제 10권 제 4호 2000년 8월 참조

권 순 학 (Soon H. Kwon)
제 10권 제 4호 2000년 8월 참조