개선된 산 클러스터링 방법
Advanced Mountain Clustering Method

이 중 우, 권 순 학, 손 세 호
Jung W. Lee, Soon H. Kwon, Seo H. Son
경북 경산시 대동 214-1
영남대학교 공과대학 전자정보공학부
Tel: 053-810-3514, 1528
Fax: 053-813-8230
E-mail: infoic@shinbiro.com

ABSTRACT : We introduce an advanced mountain clustering method which uses a normalized data space, a gaussian type mountain function and a deconstruction method using mountain slope. This 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. Computational examples are presented to show the validity of the advanced mountain method.

1. 서론

일반적으로 많이 사용되는 퍼지 클러스터링 방법에는 Beztek의 FCM(Fuzzy C-Mean)[1]과 그의 단점을 보완하기 위한 개선된 방법들[2,3]이 있으나 이러한 방법들은 초기치의 설정이 어렵고 적절하지 못한 초기치의 설정은 알고리즘이 국소해에 빠져 부적절한 클러스터링 결과를 낳을 수도 있다. 또한 선험적(heuristic) 접근법으로는 Yager와 Filev의 Mountain Method(MM)[4]와 이를 보완한 Subtractive 클러스터링[5]이 있으나 결정해야 할 매개변수가 많으며 클러스터 중심 주위에서 여러개의 중심이 생기는 등으로 인해 좋은 결과를 얻기가 용이하지 않다 .

본 논문에서는 우선 MM 알고리즘을 간략히 살펴본 후, 이의 단점을 개선한 보다 효율적이며 직관적인 클러스터 중심을 찾을 수 있는 Advanced Mountain Method(AMM) 알고리즘을 제시하고 이의 타당성을 모의 실험을 통해 보인다.

2. MM의 개요

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

$$I_j = [\min_k(x_{kj}), \max_k(x_{kj})] \quad (1)$$

단, $j=(1,s), k=(1,n)$

I_j를 r_j개의 구간으로 나누고 그 격자 $\{X_1^{(j)}, \cdots, X_{rj}^{(j)}\}$가 만나는 점을 노드 (N_i)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 산높이 $M(N_i)$를 구한다.

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

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

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

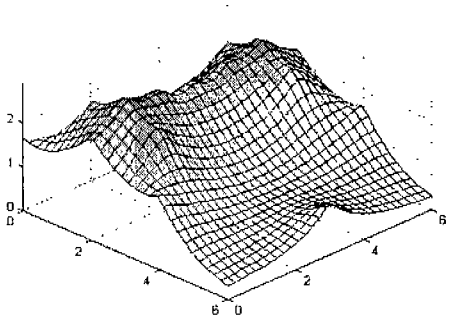
다시 그 노드의 산봉우리를 무너뜨리는 식(5) 과정을 반복하면서 중심을 찾아 나간다.

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

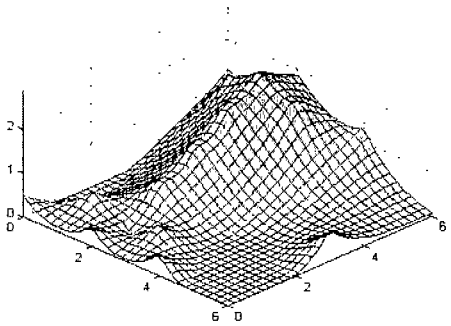
이러한 과정을 최대 산높이에 비하여 충분히 산높이가 작아질 때 까지 반복한다.

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

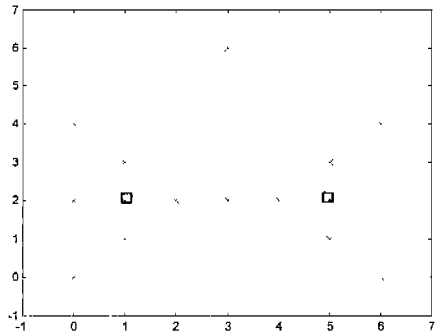
결국, 산을 만들기 위해 α, 봉우리를 붕괴하기 위해 β, 종료 조건으로 δ의 3개의 매개변수값의 선정이 필요하다.



(a) MM에 의한 산모양(α=1)

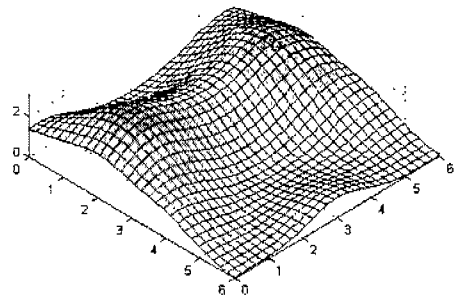


(b) 봉우리 붕괴 후(β=0.4)

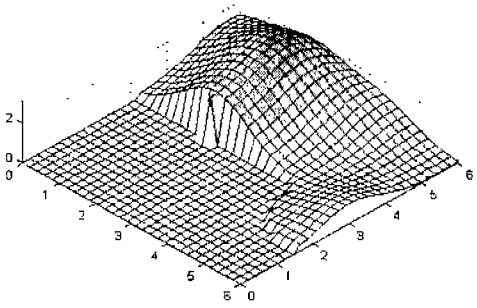


(c) 클러스터 중심(δ=0.2)

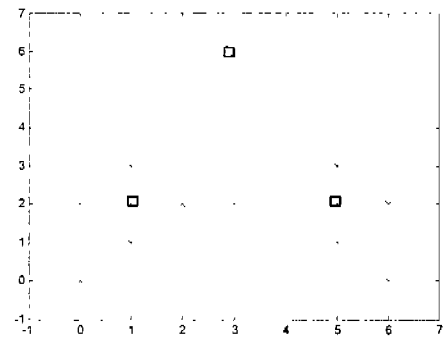
그림1. Yager's MM



(a) AMM에 의한 산모양(ω=0.1)



(b) 봉우리 붕괴 후



(c) 클러스터 중심

그림 2. 제안된 AMM

3. AMM의 개요

s차원 공간 R^s상의 n개의 데이터를 포함하는 최소공간 I_j를 식(1)과 동일하게 구성한후 r_j개의 구간으로 나눈다.

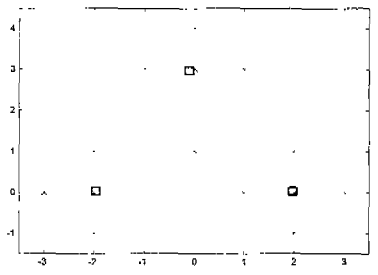
노드 N_i와 데이터 x_k의 거리를 표준화(Normalized)된 Euclidean Norm으로 구한다.

$$D_j = \max_k(x_{kj}) - \min_k(x_{kj}) \tag{7}$$

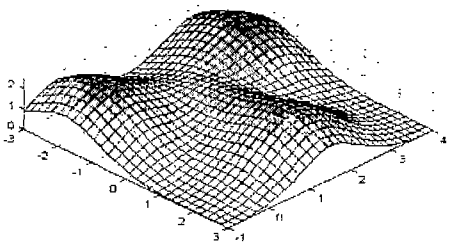
$$d_{\max} = (\Sigma_{j=1}^{s} D_j^2)^{\frac{1}{2}} \tag{8}$$

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

식(3)의 지수함수는 근접한 데이터에 대하여 산봉우리의 높이가 크게 차이가 나서 근접한

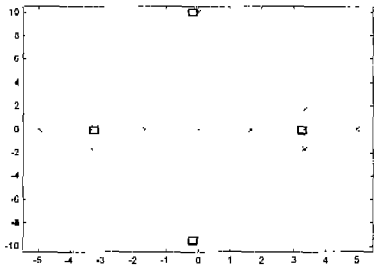


(a) 모의실험 자료

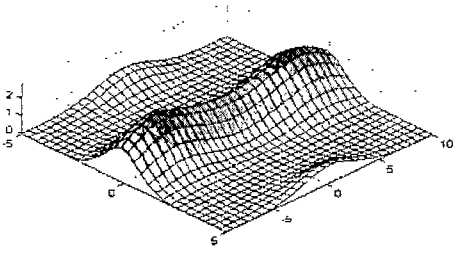


(b) AMM에 의한 산모양(ω=0.07-0.11)

그림 3. 예제1

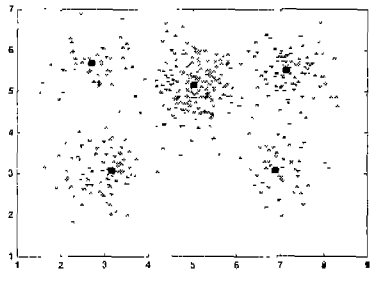


(a) 모의실험 자료

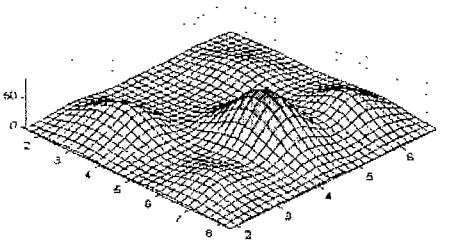


(b) AMM에 의한 산모양(ω=0.05-0.07)

그림 4. 예제2



(a) 모의실험 자료



(b) AMM에 의한 산모양(ω=0.04-0.07)

그림 5. 예제3

데이터를 동일한 클러스터로 취급하는 클러스터링의 본래의 취지에 벗어나므로 근접한 데이터에 대하여 비슷한 산봉우리를 형성하기 위하여 가우시안 함수를 적용하여 표준화된 거리로 가우시안 함수에 의하여 노드의 산높이 $M(N_i)$를 구한다.

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

산높이가 최대가 되는 점의 좌표를 식(4) $M_1^* = Max_i[M(N_i)]$에 의하여 첫 번째 클러스터 중심으로 취한후 산높이가 다시 높아지거나 바닥($M(N_i)=0$)이되는 노드를 만날 때까지 반복 하면서 산을 내려온다.

이러한 과정을 모든 노드에서 산높이 $M(N_i)$가 영이 될 때까지 반복한다.

위의 AMM알고리즘을 요약하면 다음과 같다.

1) 식(1)을 이용하여 구간 I_j를계산한다.
2) 구간 I_j를 나누어 격자를 형성한다.
3) 식(10)을 이용하여 산높이를 계산한다.
4) 산높이가 최대인 점에서 클러스터 중심을 구한후, 산높이가 다시 높아지거나 바닥

($M(N_i)=0$)이되는 노드를 만날 때까지 산을 내려오면서 산봉우리를 붕괴한다.

5) 모든 노드의 산높이가 영이 될 때까지 과정 4)를 반복한다.

AMM은 위의 알고리즘을 통해 알 수 있듯이 산봉우리를 만들 때 식(10)의 w 하나만을 조절하여 적절한 중심을 구할 수 있으므로 유용할 뿐만 아니라 표준화된 거리 $d_n(x_k, N_i)$를 사용하므로 $0<w<1$이 되고, 동일한 패턴의 확대 및 축소(Scaling)된 데이터에 대하여도 동일한 w값을 가진다.

이러한 사실들은 그림1 및 그림2를 보면 확연히 알 수 있다. Yager와 Filev이 제시한 MM에 의한 산봉우리(그림1(a))는 그 모양이 볼록(Convex)하지 않으며, 봉우리가 붕괴(Deconstruction)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 찾기위해 α, β, γ 세개의 매개변수를 조정하여야 하므로 지나치게 많은 반복수행을 필요로 함을 알 수 있다. 그러나 본 논문에서 제시하는 AMM에 의한 산봉우리(그림2(a))는 볼록한 봉우리를 형성할 가능성이 높으며, 1차로 발견된 클러스터 중심의 주위에 2차 봉우리를 형성하지 않으며 또한 조정하여야 할 매개변수가 오직 w 한 개 뿐이므로 조정이 간편하고 그 범위가 넓어(실제로 위의 데이터의 경우 $w=0.08\sim0.14$에서 적절한 중심을 찾아냄) 컴퓨터에 의한 자동클러스터링을 위한 보다 적은 반복에 의한 클러스터링 알고리즘의 가능성을 보여주고 있다.

그림3 및 그림4는 규칙적인 형태를 가지는 인공적 자료에 대하여 AMM이 적절한 중심을 찾아내는 것을 보이고 있다. 그림5는 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 랜덤자료에 대한 알고리즘의 타당성을 보이기 위한 모의 실험을 하였다. AMM이 근사적인 클러스터 중심 (2.78,5.66),(3.05,3.09),(5.12,5.15),(6.92,3.19),(7.20,5.56)을 찾아 내는 것을 보여주고 있다.

4. 결론

본 논문에서 제시한 개선된 산 클러스터링 방법(ANN)은 Yager등의 논문[4]의 결론부에서 제시한 매개변수 α, β, γ의 선택에 의한 클러스터링 결과의 민감성에 대한 문제를 매개변수를 w 한 개로 줄이므로써 부분적으로 해결하였다. 또한 노드와 데이터의 거리를 표준화 함으로써 w의 선택을 더욱 용이하게 하였다. 그리고 산봉우리 형성함수와 산봉우리 붕괴 알고리즘을 개선함으로써 클러스터 중심을 더욱 우수한 성능으로 보다 적은 반복에 의한 클러스터링이 가능하게 됨을 보였다.

향후 과제로는 모의 실험에서 보인 최적의 클러스터링된 결과와 기존의 Cluster validity index[6]를 사용한 결과와의 비교 검토를 통하여 매개변수 w와의 관계를 규명하여 보다 적은 반복에 의한 클러스터링 알고리즘에 관한 연구이다.

5. 참고문헌

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, NewYork, 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] S. Chiu, "Fuzzy Model Identification Based on Cluster Estimation," Journal of Intelligent & Fuzzy Systems, Vol.2, No.3, Sept. 1994.

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