개선된 산 클러스터링 방법의 매개변수 설정법

Identification of the Advanced Mountain Method's parameter

손 세 호, 권 순 학, 이 중 우
Seo H. Son, Soon H. Kwon, Jung W. Lee
경북 경산시 대동 214-1 영남대학교 전자정보공학부
Phone: 053-810-3514, 1528
Fax: 053-813-8230
E-mail: m0040306@chunma.yeungnam.ac.kr

Abstract : In this paper, we introduce an algorithm for identification of the Advanced Mountain Method's parameter. It consists of two phases: Phase I and Phase II. In Phase I, a given data space is divided into subspaces based on the density of the given data. In Phase II, we obtain the AMM's parameter ω by selecting the minimum of variances of subspaces obtained in Phase I. Numerical examples are presented to show the validity of the proposed method.

1. 서 론

일반적으로 주어진 자료 내의 유사한 특성을 가지는 클러스터를 분류하는 클러스터링의 방법에는 Bezdek의 FCM(Fuzzy C-Mean)[1], FCM의 단점을 보완한 방법들[2,3]이 있으나, 이러한 방법은 초기 값의 설정이 클러스터링 과정과 클러스터의 타당성에 결정적인 영향을 미친다. 또한 Yager와 Filev의 Mountain Method(MM)[4]와 이를 보완한 Subtractive 클러스터링[5]이 있으나 이 역시 결정하여야 할 매개변수가 많은 문제점이 있다. 개선된 산 클러스터링 방법(AMM)[6]은 가우시안 함수와 산의 기울기를 이용함으로써 단 하나의 매개변수(ω)만을 설정하는 알고리즘으로 개선하였으나, 가장 중요한 매개변수인 ω값을 인위적으로 설정해야 하는 문제점을 갖고 있다.

본 논문에서는 자료들이 가지고 있는 성질(밀도[7], 분산 등)을 이용하여 AMM의 매개변수(ω)를 설정하는 알고리즘을 제시하고 모의실험을 통해 제안된 알고리즘의 타당성을 보이고자 한다.

2. AMM의 개요[6]

AMM 알고리즘을 살펴보면, 주어진 s차원 공간 R^s상의 n개의 자료, $\{x_1, \cdots, x_n\}$를 포함하는 최소공간 I_j를 구성한 후 r_j개의 구간으로 나눈다. 격자 r_j가 만나는 점을 노드 N_i라 하고 각 노드와 자료 x_k의 거리를 정규화된 Euclidean Norm으로 구한다

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

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

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

$$d_n(x_k, N_i) = \frac{||x_k - N_i||_2}{d_{\max}} \quad (2)$$

(단, j=(1,...,s), k=(1,...,n))

Yager등이 제안한 MM은 산을 형성하기 위해 지수함수를 사용한다. 이러한 MM은 산을 형성

하기 위한 α, 산을 붕괴하기 위한 β, 종료조건으로 δ를 필요로 하게 된다. 하지만, AMM은 산을 형성하기 위하여 정규화된 거리를 이용한 가우시안 함수를 사용하여 노드의 산 높이 $M(N_i)$를 구한다.

$$M(N_i) = \Sigma_{k=1}^{n} e^{-d_n(x_k, N_i)^2/2w^2} \quad (3)$$

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

식(4)에 의해 산 높이가 최대가 되는 노드를 첫 번째 클러스터 중심으로 선택한 후 주변의 산 높이와 비교하여 산 높이가 다시 증가하거나 $M(N_i)=0$이 되는 노드를 만날 때까지 반복하면서 산을 붕괴시킨다. AMM은 가우시안 함수와 산의 기울기를 이용하여 MM의 매개변수(α, β, δ)를 단 하나의 매개변수(ω)를 사용하는 알고리즘으로 개선하였다. AMM 알고리즘은 다음과 같다.

(i) 식(1)을 이용하여 구간 I_j를 계산한다.

(ii) 구간 I_j를 r_j로 나누어 격자를 형성한다.

(iii) 식(3)을 이용하여 산 높이를 계산한다.

(iv) 식(4)를 이용하여 클러스터 중심을 구한 후, 산 높이가 다시 증가하거나 영이 되는 노드를 만날 때까지 산을 내려오면서 산을 붕괴시킨다.

(v) 모든 노드의 산 높이가 영이 될 때까지 (iv)를 반복한다.

3. 적절한 ω의 설정 방법

2절에서 언급하였듯이 AMM은 MM의 세 개의 매개변수(α, β, δ) 설정 문제를 개선하였다. 하지만 여전히 적절한 매개변수(ω)를 설정하여야 하는 문제점을 가지고 있다. 본 논문에서는 자료들이 가지고 있는 특성(밀도[7], 분산 등)을 이용하여 적절한 ω를 설정하는 알고리즘을 제시하고자 한다.

주어진 자료들을 정규화시킨후 자료들이 가지는 특성(자료의 밀도)을 바탕으로 공간을 분할한 후 분할된 각 공간에서의 대표값과 분산을 구한다. 분산은 주어진 자료들이 대표값, 즉 클러스터 중심으로부터 얼마나 떨어져있는가를 나타낸다. 다시 말해, 각 분산 값은 각 클러스터의 크기를 결정하는 중요한 요인이 된다. AMM에서는 산을 만들기 위해 정규화된 거리를 이용한 가우시안 함수를 사용하므로 분산과 다음과 같은 성질이 있다. 분산이 작으면 작을수록 클러스터의 크기는 작아지며 분산이 커지면 커질수록 클러스터의 크기는 커지게 된다. 즉, 분할된 공간의 분산 중 가장 작은 값을 사용하여 분류하고자 하는 최소의 클러스터의 크기를 결정할 수 있다.

이와 같은 특성을 바탕으로 구성한 매개변수 설정법은 다음과 같이 Phase I과 Phase II로 구성된다.

Phase I [7]

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

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

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

$$n_k = \max_s (x_{sj}) - \min_s (x_{sj}),$$

(단, s=1,2, j=1,2,...,n)

(ii) 정규화된 자료 $x_{j,normalized}$를 각 좌표 축 x_s상(s=1,2)에 투영시킨다.

(iii) 각각의 축 상에 투영된 각 점에서의 이웃하는 점까지의 최소 거리를 구한 후, 그 중 최대값을 구한다.

$$\Delta d^s = \max_j \left[\min_k \| x_{sj} - x_{sk} \| \right]$$

(단, s=1,2, j, k=1, ..., n, $j \neq k$)

(iv) (iii)에서 얻어진 값 Δd^s 에 적절한 상수 M를 곱한 $M \times \Delta d^s = \Delta d^s_M$을 이용하여 각 좌표계를 다음과 같이 분할한다.

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

($p \geq 0$인 정수, $x_{s_{min}}$는 축 x_s, s=1,2 상에 투영된 좌표값의 최소값이다.)

(v) 분할된 각 구간에서의 밀도를 바탕으로 밀도함수 곡선을 구한 후, 밀도함수 곡선에 존재하는 산 모양의 정점 수를 구한다.

(vi) M값을 변화시켜 산 모양의 정점의 수가 1이 될 때까지 (iv), (v)를 반복한다.

(vii) M값의 변화여도 정점의 수가 변하지 않고

가장 오래 동안 유지되는 정점의 수를 근사적 클러스터의 수로 결정한다.

Phase II

(viii) (vii)에서 결정된 클러스터를 바탕으로 공간 분할한 후 본래의 차원으로 확장한다.

(ix) (viii)에서 생성된 각 부분공간에서의 자료 $x_i \in x_j$, i=1,2, ..., n′ 에 대한 대표값과 분산을 구한다.

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

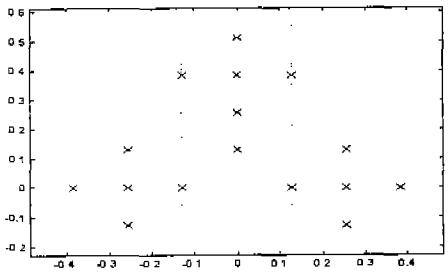
(n′>1, k=1,2,...이며 공간 분할선에 존재하는 자료는 두 공간 모두에 소속시킨다.)

(x) ω를 선택하여 AMM을 수행한다.

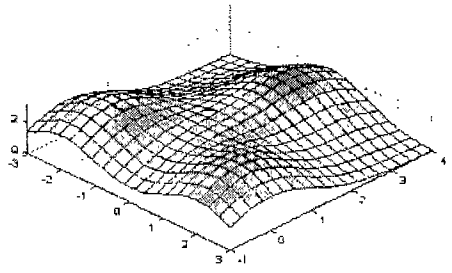
$$\omega = \min_k (\sigma_k)^{1/2} \ , \ k=1,2,...$$

4. 모의 실험 및 결과

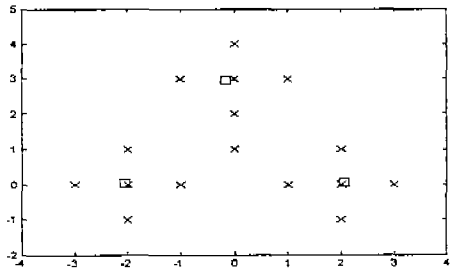
이 절에서는 본 논문에서 제시한 알고리즘의 타당성을 보이기 위해 규칙적인 형태의 자료(그림 1 및 2)와 그림 3과 같이 임의의 다섯 점(2.72,5.62),(2.95,3.09),(5.02,5.13),(7.04, 3.07), (7.18,5.41) 주위에 분산 0.5로 분포한 500개의 랜덤자료[7] 대하여 모의 실험을 수행하였다.



(a) 밀도에 의해 분할된 공간



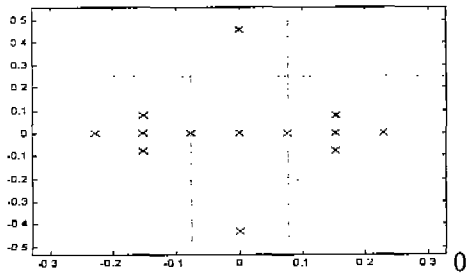
(b) AMM에 의한 산 모양 (ω=0.1145)



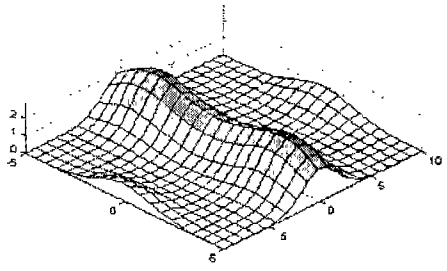
(c) 모의실험 결과

그림 1. 규칙적인 형태의 자료(I)

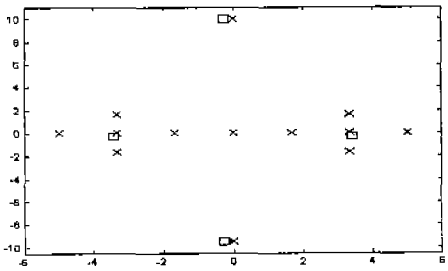
그림 1의 자료는 Phase I에 의해 6개의 부분공간(그림 1.(a))으로 나누어진다. Phase II에서 분산의 최소값은 σ_{min}^2=0.0131이다. 그림 1.(b),(c)는 설정된 $\omega=\sigma_{min}$=0.1145에 의한 AMM이 적절한 클러스터 중심을 찾아냄을 보여준다.



(a) 밀도에 의해 분할된 공간



(b) AMM에 의한 산모양(ω=0.0622)

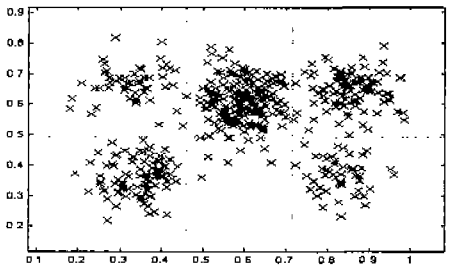


(c) 모의실험 결과

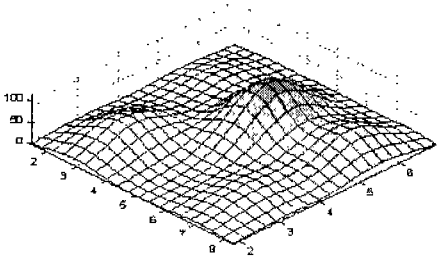
그림 2. 규칙적인 형태의 자료(II)

그림 2의 자료는 Phase I, II에 의해 주어진

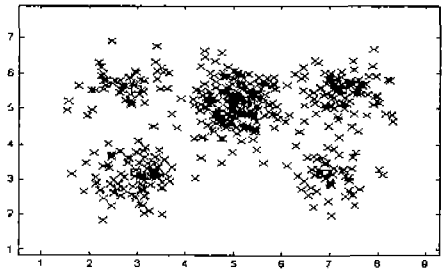
자료를 9개의 부분공간(그림 2.(a))으로 분할되며 부분공간의 분산 중 최소값은 σ_{min}^2=0.0039이 된다. 그림 2.(b),(c)는 Phase II에서 설정한 $\omega=\sigma_{min}$=0.0622로 AMM을 수행한 결과이다.



(a) 밀도에 의해 분할된 공간



(b) AMM에 의한 산 모양(ω=0.0702)



(c) 모의실험 결과

그림 3. 불규칙적인 형태의 자료(III)

그림 3의 불규칙적인 형태의 자료(III)는 Phase I에 의해 6개의 부분공간(그림 3.(a))으로 분할된다. Phase II에서 ω는 각각의 부분공간의 분산 중 최소값 σ_{min}^2=0.0049의 제곱근 0.0702이다. 그림 3.(b),(c)는 Phase II를 통해 설정한 ω를 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.

5. 결 론

본 논문에서 제시한 AMM의 매개변수 설정법은 AMM에서 사용한 가우시안 함수와 각 부분공간들의 분산과의 관계를 이용하였다. 이러한 매개변수 설정법을 통해 얻어진 ω을 사용한 AMM이 적절한 클러스터 중심을 찾아내는 것을 모의 실험을 통해 보였다.

향후 과제로는 모의 실험에서 보인 클러스터링 결과와 기존의 Cluster validity index[8]를 사용한 결과와의 비교 검토를 통한 클러스터의 타당성 검증이 이루어져야 하며, 또한 부분 공간들의 분산과 매개변수 ω와의 연관성에 대한 이론적인 고찰이 필요하며, 실제 자료에 대한 응용에 대한 연구가 필요하다.

참고문헌

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

[6] 이 중우, 권 순학, 손 세호, "개선된 산 클러스터링 방법," 2000 한국 퍼지 및 지능 시스템 학회 추계학술대회 논문집, 2000.

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

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