내부클러스터를 이용한 개선된 FCM 알고리즘에 대한 연구

안 강 식[†]·조 석 제[††]

요 약

본 논문에서는 FCM 알고리즘과 평균내부거리를 적용한 퍼지 클러스터링 알고리즘의 문제점을 해결하기 위하여 개선된 FCM 알고리즘을 제안한다. 개선된 FCM 알고리즘은 내부클러스터를 이용하여 클러스터 크기가 다른 경우에도 크기가 작은 클러스터에 일정한 소속정도를 부여할 수 있다. 그리고 이에 맞는 목적함수를 설계하고 검증한 후 데이터 분류에 사용하기 때문에 목적함수의 수렴성 문제를 극복할 수 있다. 그러므로 클러스터 크기가 다른 경우에 발생하는 FCM 알고리즘의 문제점과 목적함수의 수렴성에 문제가 있는 평균내부거리를 적용한 퍼지 클러스터링 알고리즘의 문제점을 해결할 수 있다. 제안한 알고리즘을 검증하기 위하여 제안한 알고리즘을 이용하여 데이터를 분류한 결과를 FCM 알고리즘, 평균내부거리를 적용한 퍼지 클러스터링 알고리즘을 이용하여 데이터를 분류한 결과와 각각 비교하였다. 실험을 통하여 제안한 알고리즘으로 데이터를 분류할 경우 분류 엔트로피에 의해 기존의 알고리즘들보다 더 좋은 결과를 나타냄을 알 수 있었다.

A Study on the Modified FCM Algorithm using Intracluster

Kang-Sik Ahn[†]·Seok-Je Cho[††]

ABSTRACT

In this paper, we propose a modified FCM (MFCM) algorithm to solve the problems of the FCM algorithm and the fuzzy clustering algorithm using an average intracluster distance (FCAID). The MFCM algorithm grants the regular grade of membership in the small size of cluster. And it clears up the convergence problem of objective function because its objective function is designed according to the grade of membership of it, verified, and used for clustering data. So, it can solve the problem of the FCM algorithm in different size of cluster and the FCAID algorithm in the convergence problem of objective function. To verify the MFCM algorithm, we compared with the result of the FCM and the FCAID algorithm in data clustering. From the experimental results, the MFCM algorithm has a good performance compared with others by classification entropy.

키워드 : FCM 알고리즘(FCM algorithm), 개선된 FCM 알고리즘(MFCM : modified FCM algorithm), 퍼지 클러스터링(fuzzy clustering), 평균 내부거리(average intracluster distance), 내부클러스터(intracluster)

1. 서 론

클러스터링은 주어진 데이터 집합을 비슷한 성질을 가지는 그룹으로 나누는 것으로 패턴인식, 영상처리 등 여러 공학 분야에서 전처리 과정으로 널리 적용되고 있다[1-3]. 이 중 퍼지(fuzzy) 클러스터링 알고리즘은 단순(hard) 클러스터링 알고리즘[4]에 비해 데이터의 경계가 명확하지 않더라도 클러스터링을 제대로 할 수 있기 때문에 많이 사용되고 있다[5].

기존의 퍼지 클러스터링 방법으로는 FCM(fuzzy C-means) 알고리즘[6, 7], 퍼지 ISODATA 클러스터링 알고리즘[8] 그리고 PCM(possibilistic C-means) 알고리즘[5] 등이 있다. FCM은 각 데이터와 특정 클러스터 중심과의 거리에 소속정도(degree of membership)를 부여하고 이 소속정도에 따라 데이터를 분류하는 알고리즘으로 데이터의 경계가 명확하지 않더라도 데이터를 소속정도에 따라 분류할 수 있기 때문에 단순 클러스터링 알고리즘의 문제점을 해결할 수 있다. 그러나 모든 클러스터에 대한 소속정도의 합이 1이 되는 확률적 제약조건(probabilistic constraint)을 이용하기 때문에 소속정도가 소속성(belonging)이나 적합성(compatibility) 등의 직관적인 개념과 항상 일치하지 않으며 클러스터 개수를 사전에 정해 주어야 하는 문제점이 있다. FCM 알고리즘에서의 소속정도 값은 해당 클러스터에만 관계 있는 것이 아니라 다른 클러스터와의 관계도 있기 때문에 정확한 클러스터의 원형(prototype)을 항상 추정할 수 있는 것은 아니다. 또한 클러스터의 크기가 서로 다를 경우 데이터를 제대로 분리할 수 없는 문제점이 있다.

퍼지 ISODATA 클러스터링 알고리즘[8]은 FCM 알고리

※ 이 논문은 2001년도 두뇌한국21사업에 의하여 지원되었음.
† 준 회 원 : 한국해양대학교 대학원 제어계측공학과
†† 종신회원 : 한국해양대학교 기계·정보공학부 교수
논문접수 : 2001년 10월 23일, 심사완료 : 2002년 1월 11일

즘에서 클러스터의 개수를 사전에 정해 주어야하는 문제점을 해결할 수 있다. 그러나 이 알고리즘은 데이터를 분류하기 위하여 클러스터 개수를 증가 혹은 감소시키기 때문에 시간이 많이 걸리는 단점이 있다.

Krishnapuram 등은 PCM 알고리즘을 제안하여 FCM 알고리즘의 확률적 제약조건을 개선하려 하였다[5]. PCM 알고리즘의 경우에는 FCM 알고리즘과는 달리 소속정도는 다른 클러스터와는 관계가 없고, 데이터와 해당 클러스터 중심간의 거리에만 의존하기 때문에 FCM 알고리즘에서와 같이 확률적 제약조건은 발생하지 않는다. 그러나, 이 알고리즘은 FCM 알고리즘을 먼저 실행하여 FCM 알고리즘에서 얻은 소속정도를 이용하여 클러스터 초기 중심값을 추정하는 복잡한 과정을 거치게 되고 클러스터의 크기를 사전에 알아야 하기 때문에 스케일 공간 필터링(scale space filtering)과 같은 전처리 과정이 필요하다[9].

Cho 등은 평균내부거리를 적용한 퍼지 클러스터링(FCAID : fuzzy clustering using average intracluster distance) 알고리즘을 이용하여 FCM 알고리즘이 가지고 있는 문제점을 해결하고자 하였다[10]. 이 알고리즘은 각 데이터로부터 평균내부거리 안쪽에 속하는 데이터 집합까지의 거리에 의해 소속정도를 부여하고 중심을 탐색하기 때문에 소속정도를 클러스터 크기에 관계없이 균일하게 부여할 수 있고 중심탐색능력도 향상시킬 수 있다. 그러나, 이 알고리즘은 목적함수(objective function)에 대한 명확한 해석이 없고 목적함수를 최소화하는 필요조건(necessity)과 충분조건(sufficiency)을 만족하지 않는다. 그러므로 목적함수를 최소화하는 필요조건과 충분조건에 따라 소속정도와 클러스터 중심 값을 선정해야 하는 문제점이 있다. 또한 소속정도에 도입한 평균내부거리 안쪽 데이터 소속정도가 부(negative)의 값을 가질 수 있기 때문에 목적함수가 제대로 수렴하지 않는 문제점이 있다.

본 논문에서는 FCM 알고리즘과 FCAID 알고리즘의 문제점을 극복할 수 있는 개선된 FCM(MFCM : modified FCM) 알고리즘을 제안하였다. MFCM 알고리즘은 데이터로부터 평균내부거리까지의 거리를 이용하여 소속정도를 부여하기 때문에 클러스터 크기가 다른 경우에도 데이터를 잘 분류할 수 있으며 중심탐색능력을 개선할 수 있다. 또한 MFCM 알고리즘에 적합한 목적함수를 설계하고 검증한 후 데이터를 분류하였기 때문에 목적함수의 수렴성 문제를 극복할 수 있다. MFCM 알고리즘의 성능을 평가하기 위하여 기존의 FCM 알고리즘과 FCAID 알고리즘과 비교하였으며 클러스터링 결과의 소속성과 적합성에 관한 적합도 함수인 분류 엔트로피(CE : classification entropy)[6]를 이용하여 MFCM 알고리즘의 분류 성능을 평가하였다.

2. 퍼지 클러스터링 알고리즘

퍼지 클러스터링 알고리즘은 하드 클러스터링 알고리즘에서 경계가 명확하지 않은 경우에 발생할 수 있는 데이터 분류 문제를 해결하기 위해 많이 사용되고 있다.

2.1 FCM 알고리즘

FCM 알고리즘은 퍼지 클러스터링 알고리즘 중에서 가장 폭넓게 사용되는 알고리즘으로 각 데이터와 특정 클러스터 중심과의 거리에 소속정도를 부여하고 이 소속정도에 따라 데이터를 분류하는 알고리즘이다[6, 7]. FCM 알고리즘의 목적함수는 식 (1)과 같다.

$$J_m(U, v) = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m \|v_i - x_j\|^2 \tag{1}$$

여기서, c는 클러스터 개수, n은 데이터 개수, m은 퍼지 정도를 나타내는 가중치이고, 모든 데이터 j에 대하여 $\sum_{i=1}^{c} u_{ij} = 1$이다. $X = \{x_1, x_2, \cdots, x_n\}$인 데이터 벡터 집합과 $v = \{v_1, v_2, \cdots, v_c\}$인 클러스터 중심들 사이의 소속정도를 $c \times n$인 행렬 $U(=u_{ij})$로 나타내면 식 (2)와 같고, 이 때의 클러스터 중심을 $v(=v_i)$로 나타내면 식 (3)과 같다.

$$u_{ij} = \left\{ \sum_{k=1}^{c} \left(\frac{\|v_i - x_j\|}{\|v_k - x_j\|} \right)^{\frac{2}{m-1}} \right\}^{-1}, \quad 1 \le i, k \le c,\ 1 \le j \le n \tag{2}$$

$$v_i = \frac{\sum_{j=0}^{n}(u_{ij})^m x_j}{\sum_{j=0}^{n}(u_{ij})^m} \tag{3}$$

식 (2)에서 u_{ij}는 j번째 데이터가 i번째 클러스터에 속하는 소속정도를 나타내고, 식 (3)에서 v_i는 i번째 클러스터 중심이다. 여기서 m이 1보다 큰 경우에 모든 i, j에 대해서 $v_i \neq x_j$를 만족한다고 가정하면 위의 식을 만족할 때만 (U, v)가 J_m의 최소화를 가능하게 한다. FCM 알고리즘은 식 (2)와 식 (3)을 반복하므로 J_m은 어떤 정해진 값으로 수렴한다.

그러나, FCM 알고리즘은 데이터로부터 클러스터 중심까지의 거리의 합을 최소화하는 조건에 의해 소속정도를 부여하므로 클러스터의 크기가 다른 경우 중심 탐색이 제대로 이루어지지 않으며 이로 인하여 데이터 분류 성능이 떨어진다.

2.2 FCAID 알고리즘

클러스터 크기가 다른 경우에 중심 탐색이 제대로 이루어지지 않는 FCM 알고리즘의 문제점을 개선하기 위해 Cho 등

은 FCAID 알고리즘을 제안하였다[10]. 이 알고리즘은 먼저 평균내부거리 안쪽에 속하는 데이터들의 집합을 내부클러스터(intracluster)라 정의한다. 그리고 각 데이터로부터 내부클러스터까지의 거리에 의존하여 소속정도를 구하여 소속정도를 부여한다. 이렇게 함으로써 클러스터의 크기에 관계없이 균일한 소속정도를 부여할 수 있다. 이 알고리즘의 목적함수는 식 (4)와 같다.

$$J_m(U, v) = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m\|x_j - v_i\|^2 - \sum_{i=1}^{c}\eta_i\sum_{j=1}^{n}(u_{ij})^m \quad (4)$$

여기서, 모든 j에 대해 $\sum_{i=1}^{c}u_{ij}=1$이고 η_i는 적당한 양의 정수로 내부클러스터 크기를 나타낸다. m이 1보다 큰 경우에 모든 i,j에 대해서 $v_i \neq x_j$를 만족한다면 J_m을 최소화하도록 하는 조건 (U, v)는 식 (5)와 식 (6)과 같다.

$$u_{ij} = \left\{\sum_{k=1}^{c}\left(\frac{\|v_i - x_j\| - \eta_i}{\|v_k - x_j\| - \eta_k}\right)^{\frac{2}{m-1}}\right\}^{-1} \quad (5)$$

$$v_i = \frac{\sum_{j=0}^{n}(u_{ij})^m x_j}{\sum_{j=0}^{n}(u_{ij})^m} \quad (6)$$

내부클러스터는 평균내부거리 안쪽에 속하는 데이터 집합으로 해당 클러스터 크기와 밀도에 관한 정보이다. 그러므로 내부클러스터 크기를 나타내는 η_i는 평균내부거리와 관계가 있다. Krishnapuram은 η_i를 식 (7)을 이용하여 추정하였다[7]. η_i는 소속정도를 먼저 구하고 소속정도의 값을 이용하여 구할 수 있다.

$$\eta_i = K\frac{\sum_{j=1}^{n}(u_{ij})^m\|x_j - v_i\|^2}{\sum_{j=1}^{n}(u_{ij})^m} \quad (7)$$

여기서 K는 보통 1보다 작은 값을 선택하는데 이는 데이터 개수가 많을 경우 η_i를 크게 설정하면 발산할 위험이 있기 때문이다. η_i를 식 (8)로도 표현할 수 있는데 여기서 Π_i는 데이터가 i번째 클러스터에 속할 소속정도의 집합을 나타내고 $(\Pi_i)_\alpha$는 적절한 수준의 α-절단(cut)이다[11].

$$\eta_i = \frac{\sum_{x_j\in(\Pi_i)_\alpha}\|x_j - v_i\|^2}{|(\Pi_i)_\alpha|} \quad (8)$$

그러나, FCAID 알고리즘은 실제 사용한 소속정도와 클러스터 중심으로 목적함수의 수렴성을 검증할 수 없으며, 식 (5)의 데이터 중심과 평균내부거리 간의 거리에 의한 소속함수의 값이 부의 값을 가질 수 있다. 그러므로 평균내부거리를 적용한 퍼지 클러스터링 알고리즘을 실제 데이터에 사용할 경우 목적함수의 수렴성을 검증하여야 하며 소속함수 값이 부의 값을 가지지 않게 보완할 필요가 있다. 그리고, K의 값을 0.5이상으로 설정할 경우 내부클러스터 크기가 너무 크게 설정되어 목적함수가 발산하는 문제점이 있다.

3. MFCM 알고리즘

FCAID 알고리즘의 문제점을 개선하기 위하여 본 논문에서는 MFCM 알고리즘을 제안하였다. MFCM 알고리즘은 내부클러스터를 이용하므로 클러스터 크기가 다른 경우에도 크기가 작은 클러스터에 일정한 소속정도를 부여할 수 있다. 그리고 이에 맞는 목적함수를 설계하고 검증한 후 데이터 분류에 사용하기 때문에 목적함수의 수렴성 문제를 해결할 수 있다. 그러므로 클러스터 크기가 다른 경우에 발생하는 FCM 알고리즘의 문제점을 해결할 수 있으며 목적함수의 수렴성에 문제가 있는 FCAID 알고리즘의 문제점을 극복할 수 있다. MFCM 알고리즘의 목적함수는 식 (9)와 같다.

$$J_m(U, v) = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m|d_{ij} - \eta_i| \quad (9)$$

여기서, d_{ij}는 $\|x_j - v_i\|^2$, η_i는 적당한 양의 정수로 내부클러스터 크기이다. m이 1보다 큰 경우에 모든 i,j에 대해서 $v_i \neq x_j$를 만족한다면 J_m을 최소화하도록 하는 조건 (U, v)는 식 (10), 식 (11)과 같다.

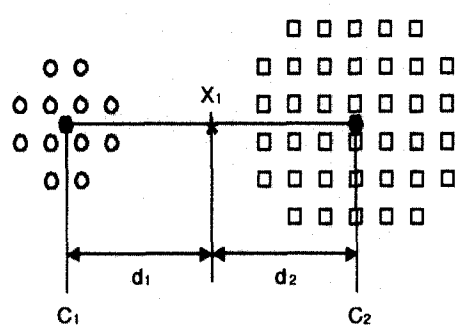
$$u_{ij} = \left\{\sum_{k=1}^{c}\left(\frac{|d_{ij} - \eta_i|}{|d_{kj} - \eta_k|}\right)^{\frac{1}{m-1}}\right\}^{-1} \quad (10)$$

$$v_i = \frac{\sum_{j=0}^{n}(u_{ij})^m x_j}{\sum_{j=0}^{n}(u_{ij})^m} \quad (11)$$

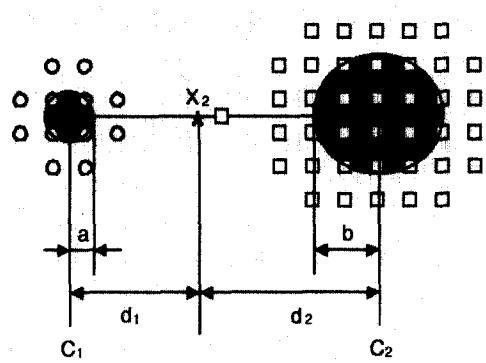
여기서, 식 (10)은 클러스터 중심과 내부클러스터 사이의 거리에 의한 소속정도가 부의 값을 가지지 않게 설정한 것으로 $\frac{|d_{ij} - \eta_i|}{|d_{kj} - \eta_k|}$에 의해서 u_{ij}는 항상 정(positive)의 값을 가지게 된다. 또한 클러스터 크기가 무한정 클 경우 각 클러스터 중심을 하나로 취급하여 데이터 분류가 제대로 되지 않기 때문에 내부클러스터 안쪽에 속하는 데이터의 소속정도를 1로 설정하지 않고 내부클러스터와 데이터 중심의 차를 소속함수에 부여하였다.

(그림 1)은 FCM 알고리즘, FCAID 알고리즘 그리고 MFCM 알고리즘의 차이를 그림으로 나타낸 것이다. C1과 C2는 각

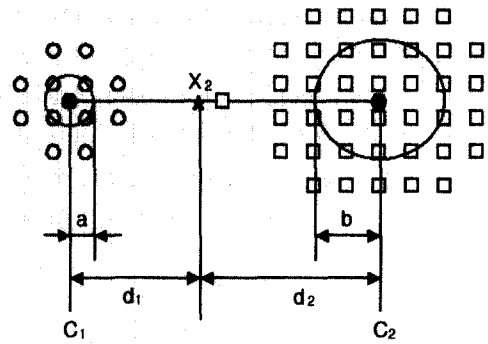
클러스터 중심, a와 b는 각 클러스터의 내부클러스터, d1과 d2는 데이터로부터 클러스터 중심까지의 거리 그리고 X1과 X2는 FCM 알고리즘, 평균내부거리를 적용한 퍼지 클러스터링 알고리즘 그리고 개선된 FCM 알고리즘에서 동일한 소속정도를 가지는 위치를 나타낸다. (그림 1)에서 X1은 직관적으로 큰 클러스터에 가깝게 위치하고 있지만 (가)의 FCM 알고리즘에서는 작은 클러스터와 큰 클러스터의 초평면(hyper-plane) 상의 한 점으로 표현한다. (나)의 평균내부거리를 적용한 퍼지 클러스터링 알고리즘에서는 내부클러스터를 고려하여 소속정도를 부여하기 때문에 X2를 작은 클러스터와 큰 클러스터의 초평면 상의 한 점으로 표현하고 (가)의 X1에 해당하는 점은 큰 클러스터에 편입시킨다. 이러한 현상은 FCM 알고리즘이 데이터로부터 클러스터 중심사이의 거리에 의해 소속정도를 부여하는 반면, FCAID 알고리즘에서는 내부클러스터 크기를 고려하여 소속정도를 부여하기 때문이다. 그러나 이 알고리즘은 내부클러스터와 클러스터 중심 사이의 거리((그림 1) (나)의 회색 부분)가 부의 영역이 될 수 있으므로 소속정도가 0과 1사이의 값이 되어야 하는 퍼지 클러스터링 알고리즘의 소속정도에 적합하지 않다. (다)의 MFCM 알고리즘에서는 식 (10)의 소속정도에 의해 내부클러스터와 클러스터 중심 사이의 거리가 부의 영역이 되지 않으며 클러스터 크기가 무한정 커지는 현상을 방지하기 위하여 내부클러스터 안쪽에 속하는 데이터 소속정도를 1로 설정하지 않았다. 그러므로, 클러스터 크기가 서로 다른 데이터 분류에서도



(가) FCM 알고리즘($d_1 = d_2$)



(나) FCAID 알고리즘($d_1 - a = d_2 - b$)



(다) MFCM 알고리즘($|d_1 - a| = |d_2 - b|$)

(그림 1) 두 개의 클러스터에서 동일한 소속정도를 가지는 각 알고리즘

FCM에서 발생하는 문제점을 해결할 수 있고 내부클러스터의 소속정도가 부의 값이 되는 것을 해결할 수 있으며 소속정도의 값을 0과 1사이의 값으로 표현할 수 있다.

3.1 MFCM 알고리즘의 수렴성

MFCM 알고리즘의 목적함수 식 (9)에 대하여 $m>1$인 경우, 모든 i, j에 대하여 목적함수를 최소화하는 최적의 짝 (U^*, v^*)를 식 (12)와 식 (13)처럼 정의한다면 (U^*, v^*)는 목적함수를 최소화하는 필요충분조건을 만족해야 한다.

$$u_{ij}^* = \frac{1}{\sum_{k=1}^{c}\left(\frac{|d_{ij}^* - \eta_i|}{|d_{kj}^* - \eta_k|}\right)^{1/(m-1)}} \tag{12}$$

$$v_i^* = \sum_{k=1}^{n}(u_{ik}^*)^m x_k \,/\, \sum_{k=1}^{n}(u_{ik}^*)^m \tag{13}$$

일반적으로 목적함수 $f(x)$를 최소화하는 x^*를 정의할 때 x^*는 아래의 최적조건을 만족해야 한다[12].

- 최적조건 1 : x^*에서 목적함수 f의 기울기 $g(x^*) = \nabla f(x^*)$는 0(zero)이다(필요조건).
- 최적조건 2 : x^*에서 목적함수 f의 헤시안(Hessian) $g(x^*) = \nabla^2 f(x^*)$는 양의 한정(positive definite)이다(충분조건).

그리고, 목적함수 $J_m(U, v)$를 최소화하는 최적의 짝 (U^*, v^*)를 구하기 위해서는 다음과 같은 세 가지 조건을 모두 고려해야 한다[8].

- 고려조건 1 : v가 고정되었다고 가정할 때, 만약 $U^* = U = [u_{ij}]$라면 U^*는 목적함수의 지역 최소값(local minimum)이다.
- 고려조건 2 : U가 고정되었다고 가정할 때, 만약 v^*

$= v = [v_1, v_2, \cdots, v_c]$라면 v^*는 목적 함수의 지역 최소값이다.

- 고려조건 3 : v와 U를 함께 고려할 때, (U^*, v^*)는 목적함수의 지역 최소값이다.

3.1.1 첫 번째 고려조건

제한된 최적화 문제에서 알고리즘의 설계와 분석에 많이 사용되는 함수는 목적함수와 제한 조건의 선형 결합으로 표현되는 Lagrange multiplier이다. v가 고정되었다고 가정하고 $u_{ij} = (w_{ij})^2$라 놓으면 식 (9)는 식 (14)와 같이 Lagrangian으로 표현된다.

$$\Phi(W, a) = \sum_{i=1}^{c}\sum_{j=1}^{n}(w_{ij})^{2m}|d_{ij} - \eta_i| + \sum_{j=1}^{n} a_j\left(\sum_{i=1}^{c} w_{ij}^2 - 1\right) \quad (14)$$

여기서, $a = (a_1, a_2, \cdots, a_n)$는 multiplier, $W = [w_{ij}^2]$는 Lagrange multiplier의 목적함수 그리고 $\Phi(W, a)$는 Lagrangian이다. 만약 (W^*, a^*)가 식 (12)를 최소화한다면, W^*와 a^*에 대한 미분은 첫 번째 최적 조건에 의해 식 (15) 및 식 (16)과 같다.

$$\frac{\partial(W^*, a^*)}{\partial a_j} = \sum_{j=1}^{c}(w_{ij}^*)^2 - 1 = 0 \quad (15)$$

$$\frac{\partial(W^*, a^*)}{\partial w_{ip}} = 2m(w_{ip}^*)^{2m-1}|d_{ip} - \eta_i| + 2(w_{ip}^*)a_p^* = 0 \quad (16)$$

식 (15)와 식 (16)을 이용하여 U^*에 대하여 정리하면 식 (12)와 같은 첫 번째 최적조건을 만족한다. 두 번째 최적 조건에 의해 식 (17)과 같이 식 (14)의 헤시안이 양의 한정이어야 한다.

$$\frac{\partial}{\partial w_{st}}\left[\frac{\partial(W^*, a^*)}{\partial w_{ip}}\right] = \begin{cases} 2m(2m-1)(w_{ip}^*)^{2m-2}|d_{ip} - \eta_i| \\ \quad + 2a_p^* \ ; \ s = i, \ t = p \\ 0; \ \text{otherwise} \end{cases} \quad (17)$$

여기서, 0이 아닌 값이 헤시안의 대각 요소이다. 식 (17)을 정리하면 식 (18)과 같이 대각 요소에만 값이 존재하는 고유치(eigenvalue)로 표현할 수 있다.

$$\lambda_p = 4m(m-1)\left[\sum_{j=1}^{c}|d_{jp} - \eta_j|^{\frac{1}{1-m}}\right]^{1-m} \quad (18)$$

여기서, $|d_{jp} - \eta_j| \geq 0$이고 $m > 1$이므로 Lagrangian의 헤시안은 양의 한정이다.

3.1.2 두 번째 고려조건

U가 고정되었다고 가정할 때, 목적함수의 지역 최소값을 구하기 위한 v^*의 첫 번째 최적조건은 식 (11)과 같다. 이 조건은 첫 번째 최적조건에 의해 v에 대한 미분으로 얻을 수 있다. 이는 임의의 방향 y에 대한 미분이 v^*에서 0이 되는 것과 같다. 즉, 클러스터 중심 v와 v^*가 같다면 임의의 방향에 대한 내적(inner product)이 0이 된다. 임의의 방향에 대한 내적을 나타내는 함수는 식 (19)와 같다.

$$\begin{aligned} h_i(t) &= \sum_{k=1}^{n}(u_{ik})^m\|x_k - (v_i^* + ty) + \sqrt{\eta_i}\|^2 \\ &= \sum_{k=1}^{n}(u_{ik})^m\langle x_k - v_i^* - ty - \sqrt{\eta_i}, x_k - v_i^* - ty - \sqrt{\eta_i}\rangle \end{aligned} \quad (19)$$

여기서, $h_i(t)$는 임의의 방향 y에서의 목적함수, t는 y의 크기를 나타내는 상수 그리고 $\langle z, z\rangle$는 $\|z\|^2$으로 내적을 나타낸다. 식 (19)를 미분하면 식 (20)과 같고, 이를 이용하여 v_i^*와 y의 내적이 0을 만족하는 식 (21), 식 (22)를 구할 수 있다.

$$\begin{aligned} \frac{dh_i(t)}{dt} &= \sum_{k=1}^{n}(u_{ik})^m(\langle -y, x_k - v_i^* - ty - \sqrt{\eta_i}\rangle \\ &\quad + \langle x_k - v_i^* - ty - \sqrt{\eta_i}, -y\rangle) \\ &= -2\left[\sum_{k=1}^{n}(u_{ik})^m\langle y, x_k - v_i^* - ty - \eta_i\rangle\right] \end{aligned} \quad (20)$$

$$\frac{dh_i(0)}{dt} = -2\left[\sum_{k=1}^{n}(u_{ik})^m\langle y, x_k - v_i^* - \eta_i\rangle\right] = 0 \quad (21)$$

$$\langle y, \sum_{k=1}^{n}(u_{ik})^m(x_k - v_i^* - \eta_i)\rangle = 0 \quad (22)$$

여기서, y는 임의의 방향에 대한 값이므로, 식 (22)를 만족하기 위해서는 두 번째 인자가 영 벡터이어야 하고, η_i는 내부 클러스터를 나타내므로 생략 가능하다. 그러므로, 식 (14)를 만족하는 첫 번째 최적조건을 만족한다. 두 번째 최적조건을 만족하기 위해서 두 번째 최적조건에 의해 식 (19)의 헤시안이 양의 한정이어야 한다. 식 (19)의 헤시안은 식 (23), 식 (24)와 같다.

$$h''(t) = \sum_{k=1}^{n}\sum_{i=1}^{c} 2(u_{ik})^m\langle y_i, y_i\rangle \quad (23)$$

$$h''(0) = 2\left[\sum_{i=1}^{c}\|y_i\|^2\sum_{k=1}^{n}(u_{ik})^m\right] \quad (24)$$

여기서, $h''(0) > 0$이므로 양의 한정이다.

3.1.3 세 번째 고려조건

v와 U를 함께 고려할 때 (U^*, v^*)는 목적함수의 지역 최소값이다. 이 경우의 첫 번째 최적조건과 두 번째 최적조건은 첫 번째 고려조건과 두 번째 고려조건을 사용한다. 세 번째 고려조건에 대한 내용은 Bezdek에 의해 증명되었다 [8, 13].

4. 실험 및 고찰

MFCM 알고리즘의 성능을 평가하기 위하여 다음과 같은 실험을 하였다. 그리고 분류 성능을 평가하기 위하여 식 (25)와 같이 분류 엔트로피의 타당성 측정함수를 사용하였다[6].

$$CE = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}[u_{ij}\log_a(u_{ij})] \qquad (25)$$

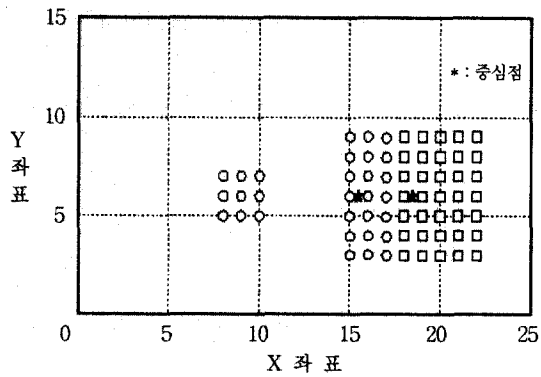
여기서 $a \in (1, \infty)$이다. 분류 엔트로피는 소속정도를 구한 후 각 데이터와 클러스터 중심의 관계를 나타내는 함수로 각 데이터가 클러스터 중심에 잘 분류될수록 작은 값을 가진다. 실험에 사용한 데이터는 크기가 다른 두 개의 클러스터로 이루어져 있는 데이터 집합 A와 크기가 같은 두 개의 클러스터로 이루어져 있는 데이터 집합 B이다.

(그림 2)는 데이터 집합 A, B를 각각 퍼지 가중치 2, 3으로 설정하여 실험한 FCM 알고리즘 결과이다. (그림 2)의 (가)와 (나)에서 데이터 집합 A를 퍼지 가중치를 2에서 3으로 증가시킬 경우 데이터의 중심이 하나의 점으로 모이는 현상이 나타나는데 이는 클러스터 크기가 다른 경우 데이터 분류를 제대로 하지 못하는 FCM 알고리즘의 문제점을 나타낸다. 그러나 (그림 2)의 (다)와 (라)에서와 같이 클러스터 크기가 같은 경우에는 이러한 현상이 일어나지 않는다.

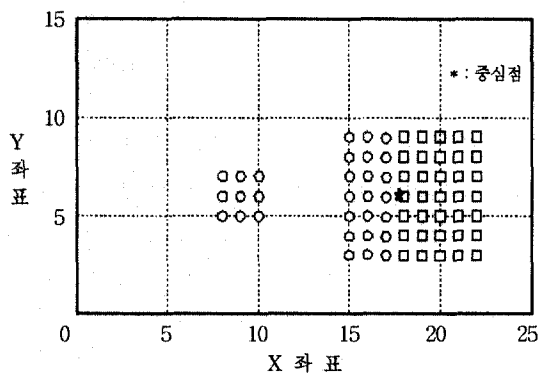
(그림 3)은 데이터 집합 A, B를 각각 퍼지 가중치 2, 3으로, K의 값을 0.5로 설정하여 실험한 FCAID 알고리즘 결과이다. (그림 3)의 (가)와 (나)에서는 FCM 알고리즘과 같이 퍼지 가중치를 2에서 3으로 증가시킬 경우에도 데이터 중심이 하나의 점으로 모이는 현상은 나타나지 않는다. 이는 FCAID 알고리즘이 클러스터 크기가 다른 경우 데이터 분류 성능이 FCM 알고리즘보다 우수하는 것을 나타낸다. (다)와 (라)는 클러스터 크기가 같을 경우의 데이터 분류 성능을 나타낸다. FCM 알고리즘과 마찬가지로 클러스터 크기가 같을 경우에는 데이터 분류 성능이 우수하였다.

(그림 4)는 데이터 집합 A, B를 각각 퍼지 가중치 2, 3으로, K의 값을 0.5로 설정하여 실험한 MFCM 알고리즘 결과이다. (그림 4)에서 (가)와 (나)는 (그림 3)에서의 (가)와 (나)보다 클러스터 중심을 더 정확하게 추정하는데 이는 퍼지 가중치가 증가하여도 FCAID 알고리즘보다 더 효과적으로 데이터를 분류한다는 것을 의미한다.

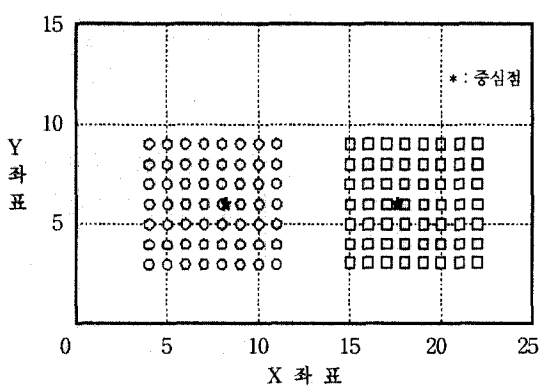
(그림 5)은 데이터 집합 A를 MFCM 알고리즘에 따라 클러스터 중심에 수렴하는 궤적을 나타낸다. 초기 클러스터 중심은 임의의 값으로 설정(여기서는 좌표 (0, 5), (20, 5)로 설정)하였으며 정지조건은 알고리즘을 2000번 반복 수행했을 때 전 단계의 소속정도와 현 단계의 소속정도 차가 최소가 되는 값을 사용하였다. 그림에서 알 수 있듯이 MFCM 알고리즘에



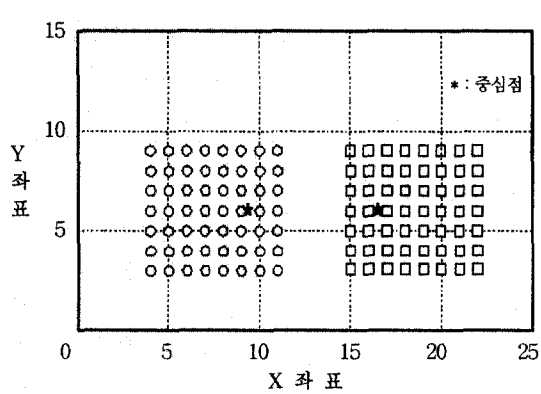
(가) 데이터 집합 A, m = 2



(나) 데이터 집합 A, m = 3

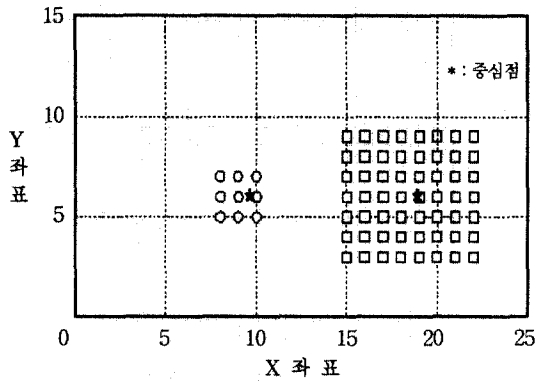


(다) 데이터 집합 B, m = 2

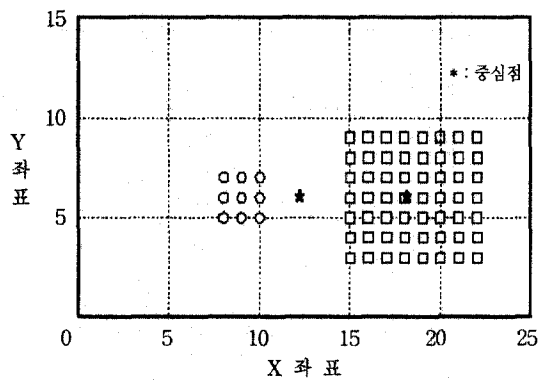


(라) 데이터 집합 B, m = 3

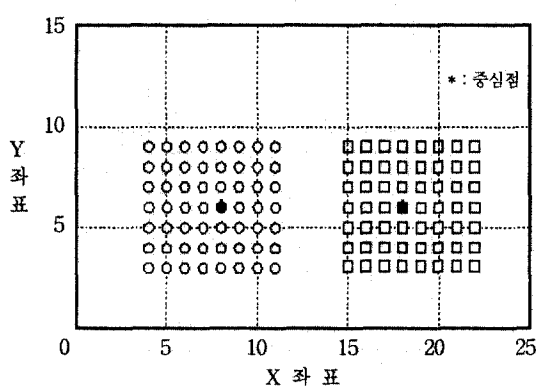
(그림 2) FCM 알고리즘을 이용한 데이터 분류



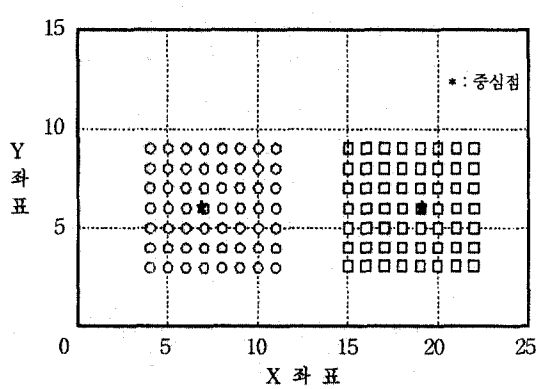
(가) 데이터 집합 A, m = 2



(나) 데이터 집합 A, m = 3

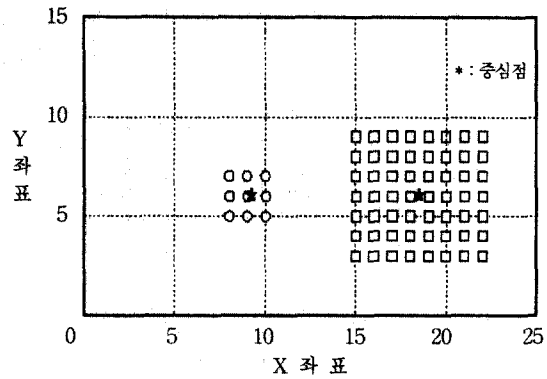


(다) 데이터 집합 B, m = 2

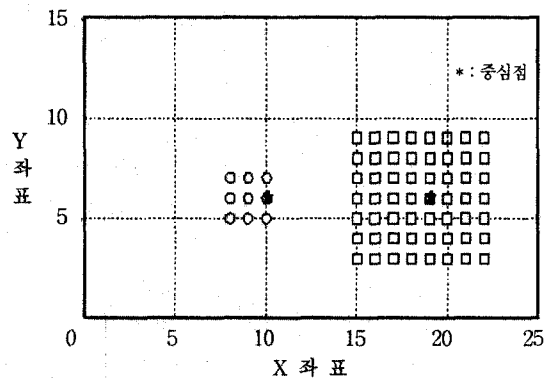


(라) 데이터 집합 B, m = 3

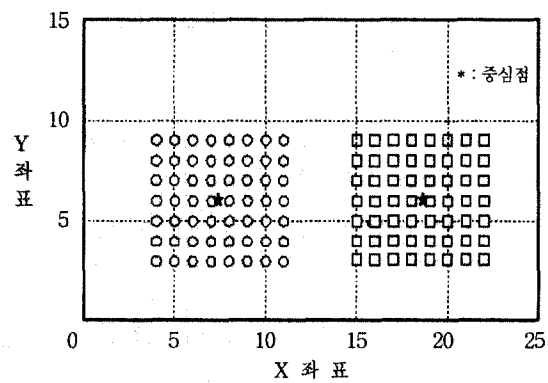
(그림 3) FCAID 알고리즘을 이용한 데이터 분류



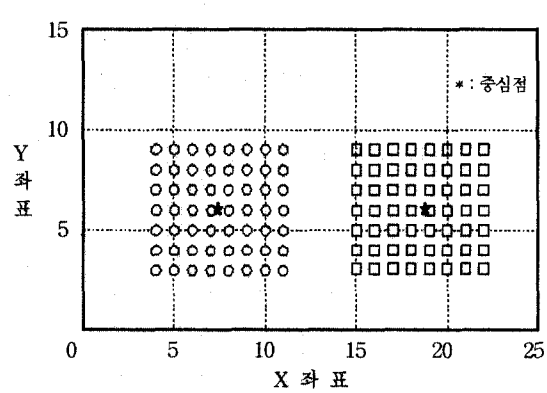
(가) 데이터 집합 A, m = 2



(나) 데이터 집합 A, m = 3

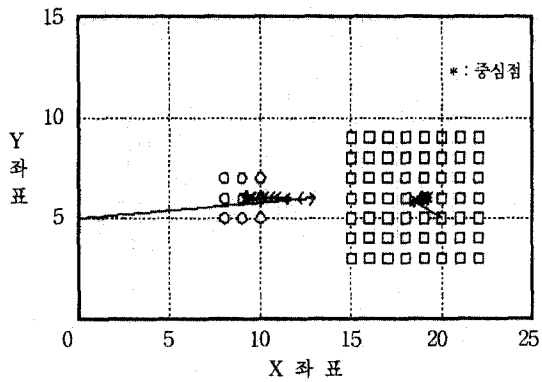


(다) 데이터 집합 B, m = 2



(라) 데이터 집합 B, m = 3

(그림 4) MFCM 알고리즘을 이용한 데이터 분류

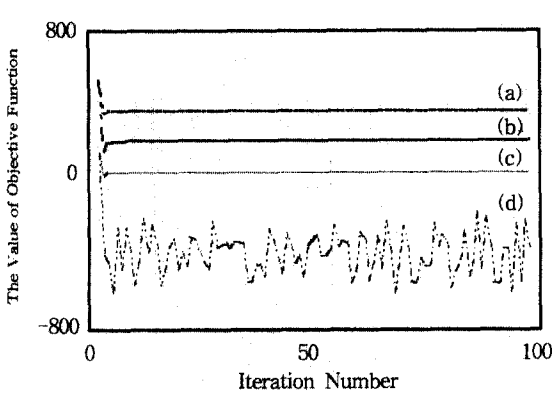


(그림 5) MFCM 알고리즘의 궤적(K = 0.1, m = 2)

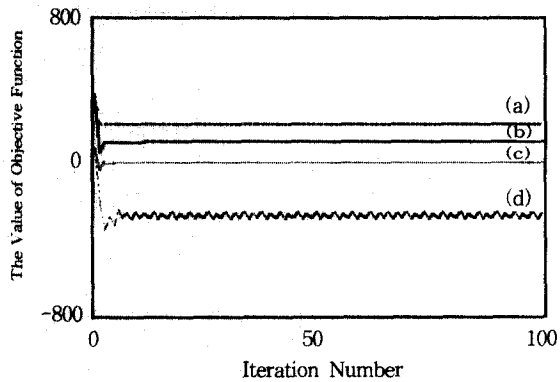
따라 클러스터 중심을 제대로 찾는다.

(그림 6)은 반복회수에 따른 목적함수의 값의 변화를 나타낸다. (가)와 (나)는 데이터 집합 A를 퍼지 가중치 2, 3으로 하였을 때의 목적함수의 값을 나타내고, (다)와 (라)는 데이터 집합 B를 퍼지 가중치 2, 3으로 하였을 때의 목적함수의 값을 나타낸다. 각 그림에서 (a)는 K의 값을 0.1로 설정한 MFCM 알고리즘, (b)는 K의 값을 0.5로 설정한 MFCM 알고리즘, (c)는 K의 값을 0.1로 설정한 FCAID 알고리즘 그리고 (d)는 K의 값을 0.5로 설정한 FCAID 알고리즘의 목적함수의 값을 나타낸다. 각 그림에서 (a)~(c)의 목적함수는 최소값으로 수렴하는 것을 알 수 있으나, (d)의 목적함수는 최소값으로 수렴하지 않는다. 이는 K의 값을 크게할 경우 내부클러스터 크기가 커져서 데이터 분류가 제대로 이루어지지 않고 목적함수가 발산하는 현상이다. 그러나 MFCM 알고리즘에서 K의 값을 0.5로 설정한 경우에는 목적함수가 발산하지 않는다. 이는 MFCM 알고리즘이 FCAID 알고리즘보다 내부클러스터 크기가 커지더라도 데이터 분류 성능이 저하되지 않으며 내부클러스터 크기에 관계없이 데이터 분류를 제대로 할 수 있음을 의미한다.

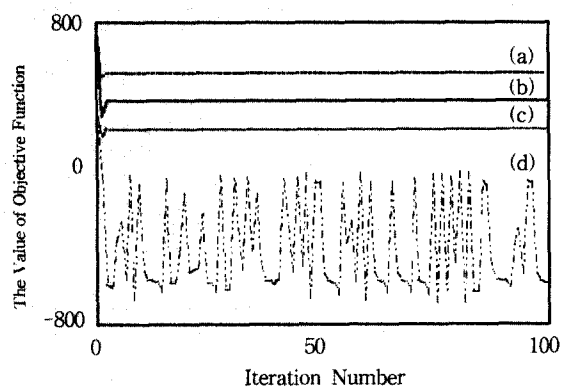
<표 1>과 <표 2>는 FCAID 알고리즘과 MFCM 알고리즘의 반복회수와 정지조건을 나타낸다. 여기서, 반복회수는 목적함수의 값이 더 이상 변하지 않을 때의 반복회수를 나타



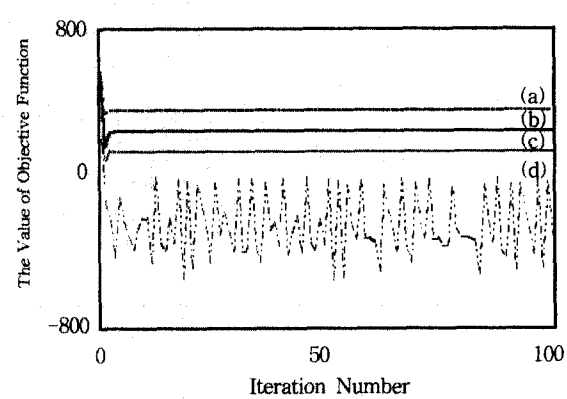
(가) 데이터 집합 A, m = 2



(나) 데이터 집합 A, m = 3



(다) 데이터 집합 B, m = 2



(라) 데이터 집합 B, m = 3

(그림 6) 반복회수에 따른 목적함수 값의 변화

(a) MFCM 알고리즘(K = 0.1)
(b) MFCM 알고리즘(K = 0.5)
(c) FCAID 알고리즘(K = 0.1)
(d) FCAID 알고리즘(K = 0.5)

내고, 정지조건은 이 때의 소속정도와 한 단계 이전의 소속정도 차를 나타낸다. <표 1>에서 K가 0.5일 때 정지조건의 값이 K가 0.1일 때보다 높은 것을 알 수 있는데 이는 목적함수 값이 최소화되는 것이 아니라 (그림 6)에서와 같이 K의 값이 증가함에 따라 목적함수 값이 발산하기 때문이다. 그러나

<표 2>의 MFCM 알고리즘에서는 K의 값이 증가하여도 정지조건의 값이 커지지 않는데 이는 알고리즘에 적합한 목적함수를 설계하였기 때문이다.

〈표 1〉 FCAID 알고리즘의 반복회수와 정지조건

K	m	데이터 집합 A		데이터 집합 B	
		반복 회수	정지 조건	반복 회수	정지 조건
0.1	2	79	4.1694e-015	38	6.5915e-015
	3	174	8.2850e-015	45	1.5871e-014
0.2	2	1250	6.8673	1149	0.2148
	3	11	13.5866	684	0.4002

〈표 2〉 MFCM 알고리즘의 반복회수와 정지조건

K	m	데이터 집합 A		데이터 집합 B	
		반복 회수	정지 조건	반복 회수	정지 조건
0.1	2	97	6.7364e-015	32	4.2434e-015
	3	103	1.5647e-014	46	0
0.2	2	65	5.1628e-015	31	5.5021e-015
	3	95	1.1706e-014	51	7.5773e-015

<표 3>은 FCAID 알고리즘과 MFCM 알고리즘의 분류 엔트로피를 나타낸다. <표 3>에서 알 수 있듯이 MFCM 알고리즘의 분류 엔트로피가 FCAID 알고리즘의 분류 엔트로피보다 더 작은 값을 가지는데 이는 MFCM 알고리즘이 FCAID 알고리즘보다 데이터 분류 성능이 우수함을 나타낸다.

〈표 3〉 각 알고리즘의 분류 엔트로피

K	m	데이터 집합 A		데이터 집합 B	
		FCAID	MFCM	FCAID	MFCM
0.1	2	0.2891	0.2116	0.2258	0.1747
	3	0.5182	0.4589	0.4663	0.4124
0.2	2	0.2270	0.2011	0.4592	0.1824
	3	0.5726	0.4790	0.4783	0.4313

5. 결 론

본 논문에서는 FCM 알고리즘과 FCAID 알고리즘의 문제점을 해결하기 위하여 MFCM 알고리즘을 제안하였다. MFCM 알고리즘은 평균내부거리 안쪽에 속하는 데이터들의 집합인 내부클러스터를 이용하므로 클러스터 크기가 다른 경우에도 크기가 작은 클러스터에 일정한 소속정도를 부여할 수 있다. 그러므로 클러스터 크기가 다른 경우 데이터를 오분류하는 FCM 알고리즘의 문제점을 해결할 수 있었다. 또한 MFCM 알고리즘에 적합한 목적함수를 설계하고 이를 검증한 후 데이터 분류에 사용하였기 때문에 FCAID 알고리즘에서와 같이 목적함수 값이 발산하지 않았다. 또한 내부클러스터 안쪽에 속하는 데이터의 소속정도를 0에서 1사이의 값으로 만들어 퍼지 소속함수에 타당한 값으로 설정하였다.

그러나 MFCM 알고리즘은 내부클러스터의 영향을 많이 받으므로 내부클러스터에 대한 추가적인 연구가 필요하고 내부클러스터 안쪽에 대한 명확한 해석이 필요하다.

참 고 문 헌

[1] T. A. Runkler and J. C. Bezdek, "Alternating Cluster Estimation : A New Tool for Clustering and Function Approximation," *IEEE Trans. Fuzzy Syst.*, Vol.7, No.4, pp. 377-393, 1999.

[2] P. R. Kersten, "Fuzzy Order Statistics and Their Application to Fuzzy Clustering," *IEEE Trans. Fuzzy Syst.*, Vol.7, No.6, pp.708-712, 1999.

[3] J. C. Bezdek and M. M. Rrivedi, "Low Level Segmentation of Aerial Images with Fuzzy Clustering," *IEEE Trans. Syst., Man, and Cybern.*, Vol.SMC-16, No.4, pp.589-598, 1986.

[4] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley Piblishing Company, 1974.

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

[6] N. Pal and J. Bezdek, "On Cluster Validity for the Fuzzy C-Means Model," *IEEE Trans. Fuzzy Syst.*, Vol.3, No.3, pp.370-379, 1995.

[7] R. Krishnapuram, H. Frigui, and O. Nasraoui, "Fuzzy and Possibilistic Shell Clustering Algorithms and their Application to Boundary Detection and Surface Approximation," *IEEE Trans. Fuzzy Syst.*, Vol.3, No.1, pp.29-60. 1995.

[8] J. C. Bezdek, "A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms," *IEEE Trans. Patt. Anal. Machine Intell.*, Vol.PAMI-2, No.1, pp.1-8, 1980.

[9] A. P. Witkin, "Scale-Space Filtering," *Processing IJCAI-83*, pp.1019-1022, 1983.

[10] H. J. You, K. S. Ahn, and S. J. Cho, "Image Segmentation Based on the Fuzzy Clustering Algorithm using Average Intracluster Distance," 정보처리학회논문지, 제7권 제9호, pp.3029-3036, 2000.

[11] 오성권, 퍼지모델및 제어이론과 프로그램, 枝多 利, 1999.

[12] R. C. Duda, R. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd Edition, John Willey & Sons, Inc., 2001.

[13] J. C. Bezdek, "Convergence Theory for Fuzzy c-Means : Counterexamples and Repairs," *IEEE Trans. Fuzzy Syst.*, September/October, 1987.

안 강 식

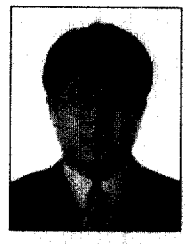
e-mail : kangsikahn@kmaritime.ac.kr

1999년 한국해양대학교 제어계측공학과 (공학사)

2001년 한국해양대학교 대학원 제어계측공학과(공학석사)

2001년~현재 한국해양대학교 대학원 제어계측공학과 박사과정

관심분야 : 색재현, 영상분할, 패턴인식, 신호처리 등



조 석 제

e-mail : sjcho@kmaritime.ac.kr

1982년 경북대학교 전자공학과(공학사)

1988년 경북대학교 대학원 전자공학과 (공학석사)

1991년 경북대학교 대학원 전자공학과 (공학박사)

1991년~현재 한국해양대학교 기계·정보공학부 부교수

관심분야 : 신호처리, 영상처리, 컴퓨터비전, 패턴인식 등