2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

내부클러스터를 이용한 개선된 FCM 알고리즘에 관한 연구

안강식*, 이동욱*, 조석제*
*한국해양대학교 제어계측공학과
joyehddnr@hanmail.net

A Study on the Modified FCM Algorithm using Intracluster

Kang-Sik Ahn*, Dong-Wook Lee*, Seok-Je Cho*
Dept. of Control & Instrumentation Engineering,
Korea Maritime University

요약

본 논문에서는 서로 다른 크기의 클러스터에 대해서 효과적으로 데이터를 분류할 수 있는 내부클러스터를 이용한 개선된 FCM 알고리즘을 제안하였다. 내부클러스터는 평균내부거리 안쪽에 속하는 데이터 집합으로 클러스터의 크기와 밀도에 비례한다. 그러므로 이를 이용한 개선된 FCM 알고리즘은 기존의 FCM 알고리즘이 클러스터 크기가 다를 경우 퍼지분할과 중심탐색을 제대로 하지 못하는 문제점을 개선할 수 있다. 실험을 통하여 개선된 FCM 알고리즘이 분류 엔트로피에 의해 기존의 FCM 알고리즘보다 더 좋은 결과를 나타냄을 알 수 있었다.

1. 서론

클러스터링은 주어진 데이터 집합을 비슷한 성질을 가지는 그룹으로 나누는 것으로 패턴인식, 영상처리 등 여러 공학 분야에 널리 적용되고 있다[1-3]. 기존의 클러스터링 방법으로는 하드 클러스터링, FCM(Fuzzy C-Means)[4-5] 그리고 PCM (Possibilistic C-Means)[6] 등이 있다. FCM 알고리즘은 데이터간의 경계가 명확하지 않을 때 영역을 제대로 분할할 수 없는 하드 클러스터링의 문제를 해결할 수 있으며, PCM 알고리즘보다 계산량이 작다. 하지만 소속정도의 합이 1이 되는 확률적 제약조건을 이용하므로 소속함수 값이 소속성 또는 적합성 등의 직관적인 개념과 항상 일치하지는 않는다. 그리고 클러스터 크기가 다른 경우 중심탐색에 있어 문제점이 발생한다.

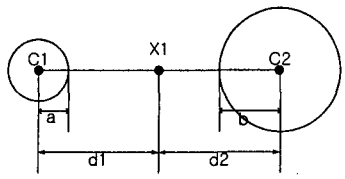
S. J. Cho 등은 평균내부거리를 이용하여 FCM 알고리즘의 문제를 해결하고자 하였다[7]. 그러나 이 방법은 PCM 알고리즘의 목적함수를 그대로 사용하였기 때문에 소속함수와 중심탐색에 문제가 있다. 그래서, 본 논문에서는 제안한 목적함수에 적합한 소속함수와 내부클러스터를 이용한 개선된 FCM 알고리즘을 제안하였다.

제안한 알고리즘은 데이터로부터 내부클러스터까지의 거리에 의해 소속정도를 부여하고 중심을 탐색한다. 내부클러스터는 평균내부거리 안에 속하는 데이터들의 집합을 말하며 클러스터 크기와 밀도에 비례한다[7]. 그러므로 클러스터 크기에 상관없이 중심탐색능력을 향상시킬 수 있었다.

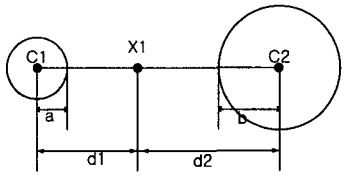
2. 평균내부거리를 이용한 개선된 FCM 알고리즘

본 논문에서는 FCM 알고리즘이 클러스터 크기가 다른 경우 데이터를 제대로 분할하지 못하는 문제를 개선하기 위해 내부클러스터를 이용한 개선된 FCM 알고리즘을 제안하였다. 먼저 평균내부거리 안쪽에 속하는 데이터들의 집합을 내부클러스터라 정의한다. 그리고 데이터로부터 내부클러스터까지의 거리에 의존하여 소속정도를 구하므로 클러스터 크기에 상관없이 균일하게 소속정도를 부여하고 클러스터 중심을 정확하게 찾을 수 있다.

(그림 1)에서 C1과 C2는 클러스터 중심, a와 b는 평균내부거리 그리고 d1과 d2는 데이터로부터 클러스터 중심까지의 거리를 나타낸다. (a)의 FCM 알고



(a) FCM 알고리즘 (d1=d2)



(b) 제안된 알고리즘 (d1-a=d2-b)

(그림 1) 두 개의 클러스터에서 동일한 소속정도를 가지는 FCM과 제안한 알고리즘

리즘에서 X1은 큰 클러스터에 가깝게 위치하고 있지만 동일한 소속정도를 부여한다. 하지만 (b)의 제안한 알고리즘에서는 내부클러스터를 고려하므로 클러스터 크기에 관계없이 균일한 소속정도를 부여할 수 있다.

제안한 알고리즘의 목적함수는 내부클러스터를 고려한 데이터로부터 클러스터 중심까지의 거리이다. 그러므로 소속정도는 중심으로부터 데이터에 이르는 분산 정도를 최소화하도록 한다.

3. 목적함수

제안한 알고리즘의 목적함수는 식 (1)과 같이 데이터 집합을 데이터로부터 내부클러스터까지의 거리를 최소로 하여 분할하도록 설계한다.

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

여기서, n은 데이터 개수, c는 클러스터 개수이고 $m\in[1,\infty]$은 퍼지정도를 나타내는 값이다. $U=[u_{ij}]$는 j번째 데이터가 i번째 클러스터에 속하는 소속정도를 나타내고 $c\times n$인 행렬로 나타나는 퍼지 분할, v는 클러스터 중심이다. 거리 $d_{ij}=\|v_i-x_j\|^2>0$ 이며, η_i는 적당한 양의 정수로 내부클러스터의 크기이다. $m>1$인 경우, 모든 i, j에 대하여 목적함수를 최소화하는 최적의 짝 (U^*, v^*)를 식 (2)와 (3)으로 정의한다면 (U^*, v^*)는 목적함수를 최소화하는 지역 최소값이 되어야 한다.

$$u_{ij}^*=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}^*-\eta_i}{d_{kj}^*-\eta_k}\right)^{1/(m-1)}} \tag{2}$$

$$v_i^*=\sum_{k=1}^{n}(u_{ik}^*)^m x_k \Big/ \sum_{k=1}^{n}(u_{ik}^*)^m \tag{3}$$

그러나, 이러한 (U^*, v^*)는 목적함수의 최소화를 만족하는 필요조건이므로 (U^*, v^*)가 목적함수를 최소화하기 위한 충분조건을 만족해야 한다. 목적함수의 최소화를 만족하는 충분조건은 다음과 같다.

충분조건 1. v가 고정되었다고 가정할 때, 만약 $U^*=U=[u_{ij}]$라면 U^*는 목적함수의 지역 최소값이다.

충분조건 2. U가 고정되었다고 가정할 때, 만약 $v^*=v=[v_1, v_2, \cdots, v_c]$라면 v^*는 목적함수의 지역 최소값이다.

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

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

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

그러므로 위의 필요조건들과 충분조건들은 이러한 최적조건을 모두 고려해야만 한다.

3. 1 충분조건 1

제한된 최적화 문제에서 알고리즘의 설계와 분석에 많이 사용되는 함수는 목적함수와 제한 조건의 선형 결합으로 표현되는 Lagrangian multiplier이다. v가 고정되었다고 가정하고 $u_{ij}=(w_{ij})^2$라 놓으면 식 (1)은 식 (4)와 같이 Lagrangian으로 표현된다.

$$\Phi(W,\alpha)=\sum_{i=1}^{c}\sum_{j=1}^{n}(w_{ij})^{2m}(d_{ij}-\eta_i)+\sum_{j=1}^{n}\alpha_j\left(\sum_{i=1}^{c}w_{ij}^2-1\right) \tag{4}$$

여기서, $\alpha=(\alpha_1, \alpha_2, \cdots, \alpha_n)$는 multiplier, $W=[w_{ij}^2]$는 Lagrange multiplier의 목적함수 그리고 $\Phi(W,\alpha)$는 Lagrangian이다. 만약 (W^*, α^*)가 식 (4)를 최소화한다면, W^*와 α^*에 대한 미분은 첫 번째 최적조건에 의해 식 (5)와 식 (6)과 같다.

$$\frac{\partial(W^*,\alpha^*)}{\partial\alpha_j}=\sum_{j=1}^{c}(w_{ij}^*)^2-1=0 \tag{5}$$

$$\frac{\partial(W^*,\alpha^*)}{\partial w_{lp}}=2m(w_{lp}^*)^{2m-1}(d_{lp}-\eta_l)+2(w_{lp}^*)\alpha_p^*=0 \tag{6}$$

식 (5)와 식 (6)을 이용하여 U^*에 대하여 정리하면 식 (2)와 같은 필요조건을 만족한다. 충분조건을 만족하기 위해서는 두 번째 최적 조건에 의해 식 (7)과 같이 식 (4)의 Hessian이 양의 한정이어야 한다.

$$\frac{\partial}{\partial w_{st}}\left[\frac{\partial(W^*, a^*)}{\partial w_{lp}}\right] = \begin{cases} 2m(2m-1)(w_{lp}^*)^{2m-2}(d_{lp}-\eta_l) \\ \quad +2a_p^* \;;\; s=l,\; t=p \\ 0; \quad \text{otherwise} \end{cases} \tag{7}$$

여기서, 0이 아닌 값이 Hessian의 대각 요소이다. 식 (7)을 정리하면 식 (8)과 같이 기저 값(eigenvalues)으로 표현할 수 있다.

$$\lambda_p = 4m(m-1)\left[\sum_{j=1}^{c}(d_{jp}-\eta_j)^{\frac{1}{1-m}}\right]^{1-m} \tag{8}$$

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

3. 2 충분조건 2

U가 고정되었다고 가정할 때, 목적함수의 지역 최소값을 구하기 위한 v^*의 필요조건은 식 (3)과 같다. 이 조건은 첫 번째 최적 조건에 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 (9)와 같다.

$$\begin{aligned} h_i(t) &= \sum_{k=1}^{n}(u_{ik})^m \|x_k-(v_i^*+ty)-\sqrt{\eta_i}\|^2 \\ &= \sum_{k=1}^{n}(u_{ik})^m \langle x_k-v_i^*-ty-\sqrt{\eta_i}, x_k-v_i^*-ty-\sqrt{\eta_i}\rangle \end{aligned} \tag{9}$$

여기서, $h_i(t)$는 임의의 방향 y에서의 목적함수, t는 y의 크기를 나타내는 상수 그리고 $\langle z,z\rangle=\|z\|^2$로 내적을 나타낸다. 식 (9)를 미분하면 식 (10)과 같고, 이를 이용하여 v_i^*와 y의 내적이 0을 만족하는 식 (11), 식 (12)를 구할 수 있다.

$$\begin{aligned} \frac{dh_i(t)}{dt} &= \sum_{k=1}^{n}(u_{ik})^m(\langle -y, x_k-v_i^*-ty-\sqrt{\eta_i}\rangle \\ &\quad +\langle x_k-v_i^*-ty-\sqrt{\eta_i}, -y\rangle) \\ &= -2\left[\sum_{k=1}^{n}(u_{ik})^m\langle y,\; x_k-v_i^*-ty-\eta_i\rangle\right] \end{aligned} \tag{10}$$

$$\frac{dh_i(0)}{dt} = -2\left[\sum_{k=1}^{n}(u_{ik})^m\langle y,\; x_k-v_i^*-\eta_i\rangle\right]=0 \tag{11}$$

$$\langle y,\; \sum_{k=1}^{n}(u_{ik})^m(x_k-v_i^*-\eta_i)\rangle=0 \tag{12}$$

여기서, y는 임의의 방향에 대한 값이므로, 식 (12)를 만족하기 위해서는 두 번째 인자가 영 벡터이어야 하고, η_i는 내부클러스터를 나타내므로 생략 가능하다. 그러므로, 식 (3)을 만족하는 필요조건을 만족한다. 충분조건을 만족하기 위해서 두 번째 최적 조건에 의해 식 (9)의 Hessian이 양의 한정이어야 한다. 식 (9)의 Hessian은 식 (13), 식 (14)와 같다.

$$h''(t) = \sum_{k=1}^{n}\sum_{i=1}^{c}2(u_{ik})^m\langle y_i,\; y_i\rangle \tag{13}$$

$$h''(0) = 2\left[\sum_{i=1}^{c}\|y_i\|^2\sum_{k=1}^{n}(u_{ik})^m\right] \tag{14}$$

여기서, $h''(0)>0$ 이므로 양의 한정이다.

3. 3 충분조건 3

v와 U를 함께 고려할 때 (U^*, v^*)는 목적함수의 지역 최소값이다. 이 경우의 필요조건과 충분조건은 충분조건 1과 충분조건 2를 사용한다. 여기서 내부 클러스터에 해당하는 η_i는 충분조건 3의 증명에 영향을 미치지 않는다. 충분조건 3에 대한 내용은 J. C. Bezdek에 의해 증명되었다[8].

4. 실험 및 고찰

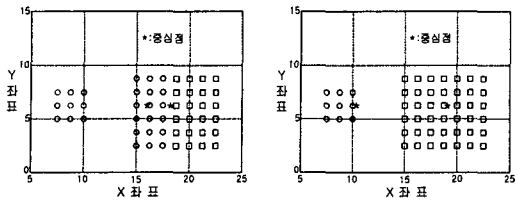
제안한 알고리즘의 성능을 평가하기 위하여 다음과 같은 실험을 하였다. 그리고 제안된 알고리즘의 분류 성능을 평가하기 위하여 식 (15)와 같이 분류 엔트로피(CE, Classification Entropy)의 타당성 측정 함수를 사용하였다[4].

$$CE = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}[u_{ij}\log_a(u_{ij})] \tag{15}$$

여기서 $a\in(1,\infty)$이고, 데이터가 잘 분류될수록 분류 엔트로피는 작은 값을 가진다.

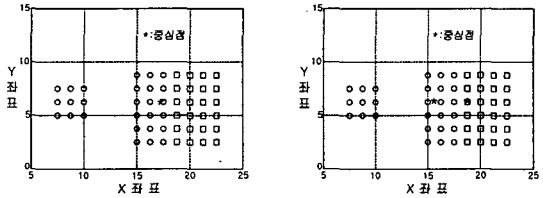
실험에 사용한 데이터 집합은 크기가 다른 두 개의 클러스터로 이루어져 있다. (그림 2)는 퍼지 가중치(m)를 2로 하여 실험한 FCM 알고리즘과 제안한 알고리즘의 클러스터링 결과이다. (a)의 FCM 알고리즘의 경우 작은 클러스터는 상대적으로 큰 클러스터의 영향을 받아 중심 값이 오른쪽으로 이동해 있다. 그러나 (b)의 제안한 알고리즘은 클러스터 크기에 따라 소속정도를 보정해 줄 수 있도록 설계되었기 때문에 상대적으로 작은 클러스터일 경우에도 중심 값을 잘 찾는다.

(그림 3)은 퍼지 가중치를 3으로 하여 실험한 FCM 알고리즘과 제안한 알고리즘의 클러스터링 결과이다. (a)의 FCM 알고리즘의 경우 두 개의 중심 값이 하나의 중심으로 나타나는데, 이는 퍼지 가중



(a) FCM 알고리즘 (b) 제안한 알고리즘

(그림 2) 퍼지 가중치가 2일 때의 데이터 중심



(a) FCM 알고리즘 (b) 제안한 알고리즘

(그림 3) 퍼지 가중치가 3일 때의 데이터 중심

치를 크게 하였을 경우의 퍼지 클러스터링의 문제점이다. 그러나 (b)의 제안한 알고리즘의 경우에는 큰 퍼지 가중치를 사용할 경우에도 FCM 알고리즘보다 더 정확한 결과를 얻을 수 있다.

<표 1>은 FCM 알고리즘과 제안한 알고리즘의 분류 엔트로피를 나타낸다. 표에서 알 수 있듯이 제안한 알고리즘에서의 분류 엔트로피가 더 작은 값을 가지는데 이는 제안한 알고리즘이 FCM 알고리즘보다 성능이 우수함을 나타낸다.

5. 결론

본 논문에서는 내부클러스터를 이용한 개선된 FCM 알고리즘을 제안하였다. 제안한 알고리즘은 평균내부거리 안쪽에 속하는 데이터들의 집합을 내부클러스터라 정의하고, 데이터로부터 내부클러스터까지의 거리를 최소로 하는 목적함수를 설계하여 클러스터 크기에 상관없이 균일한 군집화를 가능하게 하였다.

기존의 FCM 알고리즘의 목적함수는 각 데이터로부터 클러스터 중심까지의 거리를 최소로 하기 때문에 클러스터 크기가 다른 경우에는 클러스터 중심을 제대로 찾지 못하고 데이터를 제대로 분류하지 못한다. 제안한 알고리즘은 이러한 문제점을 해결할 뿐만 아니라, 크기가 작은 클러스터의 중심도 제대로 탐색할 수 있다. 그러나 제안한 알고리즘은 내부클러스터의 영향을 많이 받으므로 내부클러스터에 대한 추가적인 연구가 필요하다.

<표 1> FCM과 제안한 알고리즘의 분류 엔트로피

	FCM 알고리즘	제안한 알고리즘
m = 2	0.3078	0.1199
m = 3	0.3466	0.3195

참고문헌

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

[2] P. R. Kersten, "Fuzzy Order Statistics and Their Application to Fuzzy Clustering," *IEEE Trans. on Fuzzy Sys.*, 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. on Fuzzy Sys., Man, and Cybert.*, Vol. SMC-16, No. 4, pp. 589-598, 1986.

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

[5] R. Krishnapuram, H. Frigui, and O. Nasraoui, "Fuzzy and Possibilistic shell Clustering algorithms and their application to boundary detection and surface approximation," *IEEE Trans. on Fuzzy Sys.*, Vol. 3, No. 1, pp. 29-60. 1995.

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

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

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