PFCM 클러스터링 기법의 개선

Improvement of the PFCM(Possibilistic Fuzzy C-Means) Clustering Method

저자
(Authors)

허경용, 최세운, 우영운
Gyeongyong Heo, Sewoon Choe, Young Woon Woo

출처
(Source)

한국정보통신학회논문지 13(1), 2009.1, 177-185 (9 pages)
Journal of the Korea Institute of Information and Communication Engineering 13(1), 2009.1, 177-185 (9 pages)

발행처
(Publisher)

한국정보통신학회
The Korea Institute of Information and Communication Engineering

URL

http://www.dbpia.co.kr/Article/NODE02252963

APA Style

허경용, 최세운, 우영운 (2009). PFCM 클러스터링 기법의 개선. 한국정보통신학회논문지, 13(1), 177-185.

PFCM 클러스터링 기법의 개선

허경용* · 최세운** · 우영운***

Improvement of the PFCM(Possibilistic Fuzzy C-Means) Clustering Method

Gyeongyong Heo* · Sewoon Choe** · Young Woon Woo***

요 약

클러스터링은 주어진 데이터 포인트들을 주어진 개수의 그룹으로 나누는 비지도 학습의 한 방법이다. 클러스터링의 방법 중 하나로 널리 알려진 퍼지 클러스터링은 하나의 포인트가 모든 클러스터에 서로 다른 정도로 소속될 수 있도록 함으로써 하나의 클러스터에만 속할 수 있도록 하는 K-means와 같은 방법에 비해 자연스러운 클러스터 형태의 유추가 가능하고, 잡음에 강한 장점이 있다. 이 논문에서는 기존의 퍼지 클러스터링 방법 중 소속도(membership)와 전형성(typicality)을 동시에 계산해 낼 수 있는 Possibilistic Fuzzy C-Means (PFCM) 방법에 Gath-Geva (GG)의 방법을 적용하여 PFCM을 개선한다. 제안한 방법은 PFCM의 장점을 그대로 가지면서도, GG의 거리 척도에 의해 클러스터들 사이의 경계를 강조함으로써 분류 목적에 적합한 소속도를 계산할 수 있으며, 전형성은 가우스 형태의 분포에서 생성된 포인트들의 분포 함수를 정확하게 모사함으로써 확률 밀도 추정의 방법으로도 사용될 수 있다. 또한 GG 방법은 Gustafson-Kessel 방법과 달리 클러스터에 포함된 포인트의 개수가 확연히 차이 나는 경우에도 정확한 결과를 얻을 수 있다. 이러한 사실들은 실험 결과를 통해 확인할 수 있다.

ABSTRACT

Cluster analysis or clustering is a kind of unsupervised learning method in which a set of data points is divided into a given number of homogeneous groups. Fuzzy clustering method, one of the most popular clustering method, allows a point to belong to all the clusters with different degrees, so produces more intuitive and natural clusters than hard clustering method does. Even more some of fuzzy clustering variants have noise-immunity. In this paper, we improved the Possibilistic Fuzzy C-Means (PFCM), which generates a membership matrix as well as a typicality matrix, using Gath-Geva (GG) method. The proposed method has a focus on the boundaries of clusters, which is different from most of the other methods having a focus on the centers of clusters. The generated membership values are suitable for the classification-type applications. As the typicality values generated from the algorithm have a similar distribution with the values of density function of Gaussian distribution, it is useful for Gaussian-type density estimation. Even more GG method can handle the clusters having different numbers of data points, which the other well-known method by Gustafson and Kessel can not. All of these points are obvious in the experimental results.

키워드

Clustering, Fuzzy Clustering, PFCM, Gustafson-Kessel Method, Gath-Geva Method

* Dept. of Computer and Information Sci. and Eng., Univ. of Florida

** Dept. of Biomedical Eng., Univ. of Florida (교신저자)

*** 동의대학교 멀티미디어공학과

접수일자 2008. 07. 04

Ⅰ. 서 론

퍼지 클러스터링은 60년대 Zadeh[1]의 퍼지 집합 이론까지 거슬러 올라간다. 이를 바탕으로 Ruspini[2]가 퍼지 분할(fuzzy partition)을 클러스터링에 소개하여 기존의 Hard C-Means (HCM) 방법[3]을 퍼지 기반의 방법으로 확장할 수 있도록 했으며, Dunn[4]이 처음으로 퍼지 클러스터링을 소개하였다. 이후 Bezdek[5]에 의해 일반적인 경우로 확장되었고 현재 Fuzzy C-Means(FCM)로 알려진 방법은 일반적으로 Bezdek의 방법을 일컫는다.

FCM은 여러 분야에서 성공적으로 적용되었지만 FCM은 하나의 포인트가 모든 클러스터에 속할 정도의 합이 1이 되도록 함으로써 때때로 직관적인 분할과는 다른 결과를 보여준다. 또한 이러한 제약 사항(constraint)으로 인해 FCM은 잡음에 민감한 특성을 갖는다. 이러한 문제를 해결하기 위해 Krishnapuram과 Keller[6]는 FCM의 제약을 제거한 Possibilistic C-Means(PCM)를 제안하였다. PCM은 기존의 확률적인 (probabilistic) 방법이 아닌 가능성 이론(possibilistic theory)을 사용하여 잡음이나 특이점(outlier)들이 어느 클러스터에도 소속되지 않도록 함으로써 잡음 민감성을 제거하였다. 하지만 PCM은 각각의 클러스터들이 서로 영향을 주지 않으므로 초기화에 민감하고 때때로 중복된 클러스터를 찾아내는 문제점이 있다[7]. 이러한 FCM의 잡음 민감성과 PCM의 중첩 클러스터 문제를 극복하기 위해 확률적 측도(probabilistic measure) 또는 소속도(membership)와 가능성 측도(possibilistic measure) 또는 전형성(typicality)을 함께 사용하는 방안이 연구되었으며[8] 그 중 하나가 Possibilistic Fuzzy C-Means(PFCM)이다. PFCM은 FCM과 PCM의 장점을 결합하여 잡음에 강하고 클러스터들이 일부 중첩된 경우에도 중심을 정확히 찾아낼 수 있다.

앞서 언급한 모든 퍼지 클러스터링 방법들은 기본적으로 유클리드 거리를 사용한다. 유클리드 거리는 클러스터들이 서로 중첩되지 않고, 구형을 이루며, 각 클러스터에 속하는 데이터 포인트들의 개수가 비슷한 경우에 효과적이다. 이를 개선하기 위해 마할라노비스(Mahalanobis) 거리를 사용하여 타원형 클러스터를 찾아낼 수 있는 방법이 Gustafson과 Kessel (GK)[9, 10]에 의해 제안되었다. 하지만 GK 방법 역시 각 클러스터들이 비슷한 정도의 포인트들을 가지는 것으로 가정하고 있다. 또 한 가지 널리 사용되는 방법은 Gath와 Geva (GG)[11]에 의해 제안된 방법으로 가우스 분포 함수에 반비례하는 값을 거리로 사용한다. GG 방법은 각 클러스터의 사전 확률을 사용하여 각 클러스터에 속하는 포인트의 개수가 달라도 안정적으로 클러스터를 찾아낼 수 있다.

이 논문에서는 소속도와 전형성을 동시에 계산할 수 있는 PFCM 알고리즘에 GG 방법을 적용하여 PFCM 알고리즘을 확장한다. GG 방법은 다른 거리 척도들이 클러스터의 분포에 중점을 두어 소속도를 계산하는 것과는 달리 클러스터와 클러스터들 사이의 경계에 더 중점을 둔다. 따라서 분류 목적으로 클러스터링을 사용하는 경우 GG 방법이 더 효과적이다. 또한 PFCM 알고리즘은 전형성도 동시에 계산해 내며, 실험 결과에서 알 수 있듯이 가우스 분포 함수와 유사한 전형성 분포를 가지므로 가우스 형태의 분포 함수를 추정하기 위한 목적으로도 사용할 수 있다. GG 방법을 사용한 결과는 기존의 PFCM 방법[8], PFCM에 GK를 사용한 방법[12]과의 비교를 통해 그 차이점을 명확히 알 수 있다.

이 논문의 구성은 다음과 같다. 먼저 2장에서는 기존 퍼지 클러스터링 방법들을 다루고 이들의 문제점을 보인다. 3장에서는 GG 방법과 GK 방법에 대해 알아보고 4장에서는 이들을 PFCM에 적용한 결과를 보인다. 결론 및 향후 연구 방향에 대해서는 5장에서 언급한다.

Ⅱ. 클러스터링

주어진 데이터 포인트를 각 클러스터에 할당은 기본적으로 각 클러스터의 중심과 데이터 포인트들 사이의 거리의 함수로 결정된다. 유클리드 거리는 가장 기본적인 척도이지만 구형이 아닌 클러스터를 나타낼 수 없는 문제점이 있다. 보다 일반적인 거리 척도인 마할라노비스 거리는 식 (1)과 같이 표현된다.

$$D_{ik}^2 = (x_k - c_i)^T \Sigma_i^{-1} (x_k - c_i) \quad (1)$$

이 때 D_{ik}^2는 k번째 데이터 포인트 xk와 i번째 클러스터의 중심 ci 사이의 거리 제곱을 나타내며 Σ_i는 i번째 클러스터의 분산 행렬을 나타낸다. 유클리드 거리는 Σ_i가 단위행렬인 마할라노비스 거리의 특별한 경우에 해당하며 이 장에서는 유클리드 거리를 가정한다.

데이터 포인트와 클러스터의 중심 사이의 거리가 주어지면 각 클러스터링 알고리즘은 이를 바탕으로 데이터 포인트들을 각 클러스터에 할당한다. N 개의 데이터 포인트가 n 차원 벡터로 표현되고 C 개의 클러스터가 존재한다고 할 때, 클러스터링은 각 데이터 포인트가 각 클러스터에 속할 정도를 나타내는 N × C 분할 행렬, U(membership) 또는 T(typicality)를 최적화하는 문제로 귀결된다. 각각의 클러스터링 방법은 각기 다른 제약 사항을 가진다. HCM에서는 U의 값이 0이나 1의 값만을 가질 수 있고, FCM에서는 U의 값으로 0과 1 사이의 값을 가질 수 있지만 각 포인트의 소속도 합은 1이 되어야 한다. PCM에서 T 값 역시 0과 1 사이의 값을 가질 수 있지만 전형성의 합이 1이 될 필요는 없다.

2.1 Fuzzy C-Means (FCM)

FCM에서 각 데이터 포인트 xk는 각 퍼지 부분집합 즉, 각 클러스터에 소속되는 정도를 가지며, 목적 함수 JFCM은 식 (2)와 같이 주어진다.

$$J_{FCM} = \sum_{i=1}^{C}\sum_{k=1}^{N} u_{ik}^{m} d_{ik}^{2} \tag{2}$$

이 때 $1 < m < \infty$은 fuziness 정도를 나타내는 상수로 일반적으로 2로 설정된다. d_{ik}^2는 k번째 데이터 포인트와 i번째 클러스터 중심 사이의 거리 제곱을 나타내며, 식 (1)의 마할라노비스 거리가 그 한 예에 해당한다. 라그랑즈(Lagrange)의 방법을 사용하여 U와 ci는 각각 식 (3), (4)와 같이 구해지고[5], 일반적으로 AO (Alternating Optimization) 방법을 통해 국부 최대값을 구한다.

$$u_{ik} = \left(\sum_{j=1}^{C} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right)^{-1} \tag{3}$$

$$c_i = \frac{\sum_{k=1}^{N} u_{ik}^{m} x_k}{\sum_{k=1}^{N} u_{ik}^{m}} \tag{4}$$

그림 1은 두 개의 가우스 분포에서 생성된 데이터를 클러스터링한 결과로, 동일한 소속도 값을 갖는 포인트들을 연결한 곡선을 나타낸 것이다. 그림 1에서 알 수 있듯이, 두 클러스터 중심에서 동일한 거리에 있는 점들의 소속도 값은 모두 동일하게 0.5로 클러스터 중심으로부터의 절대적인 거리와는 무관함을 알 수 있다. 이러한 현상은 FCM이 소속도의 합이 1이 되도록 요구하기 때문으로 FCM의 단점 중 하나로 비판되고 있다. 하지만, 잡음에 의한 클러스터 중심의 편향을 효과적으로 억제하는 경우 분류를 위한 용도로는 유용하게 사용될 수 있다는 점도 간과할 수는 없다.

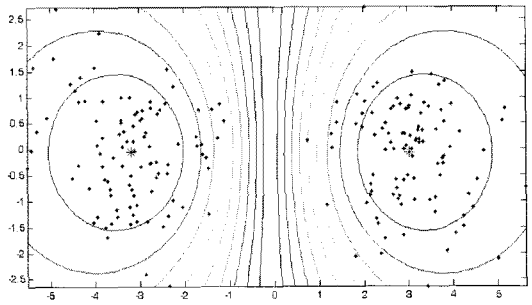

그림 1. FCM 알고리즘에 의한 클러스터링 결과
Fig. 1. Clustering result using FCM

2.2 Possibilistic C-Means (PCM)

PCM은 FCM의 직관적이지 못한 소속도 값을 개선하기 위해 소속도 값의 합이 1이 되는 제약 사항을 제거한 것이다. PCM에서는 각 포인트가 클러스터에 속하는 정도를 표현하기 위해 소속도가 아닌 전형성(typicality)을 사용한다. PCM의 목적 함수는 식 (5)과 같으며, 식에서 두 번째 항은 모든 tik 값이 0이 되는 경우 목적 함수가 최소화되는 자명해(trivial solution)를 제거하기 위해 첨가된 항이다.

$$J_{PCM} = \sum_{i=1}^{C}\sum_{k=1}^{N} t_{ik}^{m} d_{ik}^{2} + \sum_{i=1}^{c} \delta_i \sum_{k=1}^{N} (1 - t_{ik})^{m} \tag{5}$$

이 때 δ_i는 각 클러스터의 부피를 나타내는 값으로, 클러스터의 크기 추정과 특이점(outlier) 판별에 영향을 미친다. PCM 역시 FCM과 같이 AO 방법을 통해 국부 최대값을 구하는 것이 일반적이며, tik 값은 식 (6)과 같이 주어지고, ci 값은 식 (4)와 동일하다[6].

$$t_{ik} = \frac{1}{1+\left(\frac{d_{ik}^2}{\delta_i}\right)^{1/(m-1)}} \quad (6)$$

식 (6)은 식 (3)이 한 데이터 포인트와 모든 클러스터 중심 사이의 거리를 고려하는 것과 달리 한 데이터 포인트와 하나의 클러스터 중심 사이의 거리만을 고려한다. 이러한 특징은 클러스터들 사이의 상관관계를 고려하지 않음으로 해서 중첩된 클러스터 문제를 유발할 수 있다[7]. 그림 2는 두 개의 가우스 분포에서 생성된 데이터 포인트를 클러스터링한 결과를 나타낸 것으로 그림 1에서와 같이 잡음에 비직관적인 소속도를 할당하는 경우는 발생하지 않음을 알 수 있다.

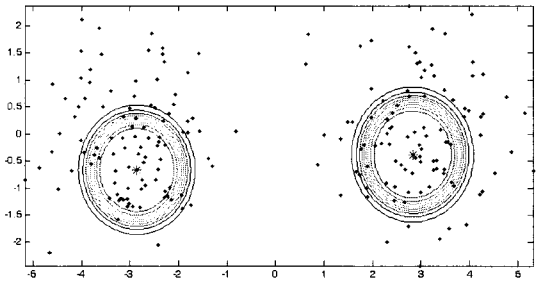

그림 2. PCM 알고리즘에 의한 클러스터링 결과
Fig. 2. Clustering result using PCM

그림 3은 인접한 2개의 가우스 분포에서 생성된 데이터 포인트들을 클러스터링 한 결과로, PCM은 1개의 클러스터 중심만을 찾아내는 반면, FCM은 2개의 클러스터 중심을 찾아내고 있다. 이러한 결과는 특히 클러스터들이 중첩되는 경우 빈번히 발생한다.

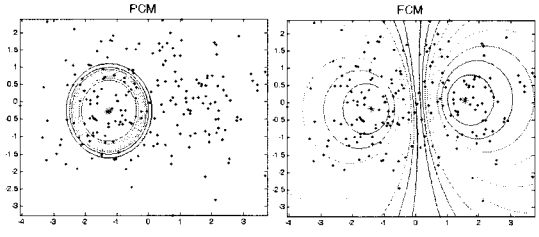

그림 3. 중첩된 클러스터에 대한 클러스터링 결과
(왼쪽 : PCM, 오른쪽 : FCM)
Fig. 3. Clustering result of overlapped clusters
(left : PCM, right : FCM)

2.3 Possibilistic Fuzzy C-Means (PFCM)

FCM과 PCM은 각각의 장점과 단점이 있으므로, 서로의 단점을 보완하기 위한 방법으로 소속도와 전형성을 함께 사용하고자 하는 시도가 있어왔고[8, 13, 14] 그 중 하나가 PFCM[8]이다. PFCM의 목적 함수는 식 (7)과 같이 주어진다.

$$J_{PFCM} = \sum_{i=1}^{C}\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^\eta)d_{ik}^2 + \sum_{i=1}^{c}\delta_i\sum_{k=1}^{N}(1-t_{ik})^\eta \quad (7)$$

이 때 a, b는 소속도와 전형성에 대한 가중치 상수이며, η는 소속도에서 m과 동일한 역할을 전형성에서 하는 상수이다. PFCM의 uik는 식 (3)과 동일하게 계산되며 tik와 ci는 각각 식 (8), (9)로 계산된다.

$$t_{ik} = \frac{1}{1+\left(\frac{bd_{ik}^2}{\delta_i}\right)^{1/(\eta-1)}} \quad (8)$$

$$c_i = \frac{\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^\eta)x_k}{\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^\eta)} \quad (9)$$

그림 4는 그림 3과 동일한 데이터를 PFCM을 통해 클러스터링한 결과이다. PFCM은 클러스터 중심의 계산을 위해 소속도와 전형성을 동시에 고려함으로써 FCM의 잡음 민감성을 완화하고 PCM의 중첩된 클러스터 문제를 해결한다.

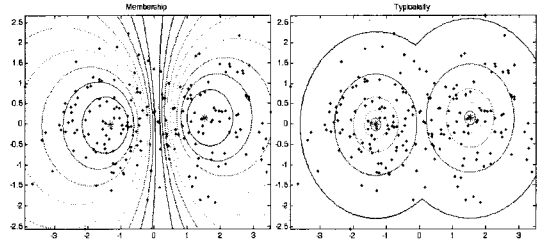

그림 4. PFCM을 이용한 클러스터링 결과
(왼쪽 : 소속도, 오른쪽 : 전형성)
Fig. 4. Clustering result using PFCM
(left : membership, right : typicality)

Ⅲ. 거리 측도 (Distance Measure)

PFCM은 FCM과 PCM의 장점을 결합한 방법이지만 기본적으로 유클리드 거리를 사용함으로 인해 몇 가지 문제점을 안고 있다. 유클리드 거리는 클러스터들이 서로 중첩되지 않고, 구형을 이루며, 각 클러스터에 속하는 데이터 포인트들의 개수가 비슷한 경우에 효과적이다. 따라서 구형이 아닌 클러스터의 경우에는 문제가 발생할 수 있다. 이를 개선하여 타원형의 클러스터를 찾아낼 수 있도록 한 방법이 Gustafson과 Kessel (GK)[9]에 의해 제안되었고, 분산 행렬의 역행렬을 구하는데 있어 특이점(singularity)이 발생하지 않도록 하는 개선된 방법[10]이 널리 사용되고 있다. 하지만 GK 방법 역시 클러스터의 형태와 크기의 다양성은 해결할 수 있지만 기본적으로 각 클러스터들이 비슷한 정도의 포인트들을 가지는 것으로 가정되고 있다. 또 한 가지 널리 사용되는 방법은 Gath와 Geva (GG)[11]에 의해 제안된 방법으로 가우스 분포 함수에 반비례하는 값을 거리로 사용한다. GG 방법은 각 클러스터의 사전 확률을 이용함으로써 각 클러스터에 속하는 포인트의 개수가 다른 경우도 다룰 수 있다.

3.1 Gustafson-Kessel 방법

Gustafson-Kessel 방법은 식 (1)에서 퍼지 분산 행렬 Σ_i를 식 (10)과 같이 계산한다.

$$\Sigma_i = \frac{\sum_{k=1}^{N} u_{ik}^m (x_k - c_i)(x_k - c_i)^T}{\sum_{k=1}^{N} u_{ik}^m} \tag{10}$$

계산된 분산 행렬은 식 (11)에서 포인트와 클러스터 중심 사이의 거리를 구하기 위해 사용된다.

$$d_{ik}^2 = (x_k - c_i)^T \left[\rho_i \cdot \det(\Sigma_i)^{1/n} \Sigma_i^{-1}\right](x_k - c_i) \tag{11}$$

이 때 ρ_i는 각 클러스터의 부피를 나타내는 파라미터이다. 식 (11)에서 Σ_i^{-1}를 구할 때, 데이터 포인트의 개수가 적은 경우 행렬식이 0의 값을 가질 수가 있다. 이러한 특이점(singularity) 발생을 방지하기 위해 식 (12)가 사용된다.

$$\Sigma_i \leftarrow (1-\gamma)\Sigma_i + \gamma \cdot \det(\Sigma_0)^{1/n} I \tag{12}$$

이 때 Σ_0는 전체 데이터 포인트를 이용해서 계산한 분산 행렬이며, γ는 가중치 상수, I는 단위행렬을 나타낸다. 계산된 분산 행렬에서 고유값(eigenvalue) λ_{ij}와 고유벡터(eigenvector) ϕ_{ij}를 계산한 후, 고유값의 최대값 $\lambda_{i,\max}$를 구하고, $\lambda_{i,\max}/\lambda_{ij} > \beta$인 고유값들을 $\lambda_{i,\max}/\beta$ 값으로 바꾼다. 즉, 고유값의 최소값을 제한함으로써 클러스터의 형태를 타원형으로 유지하도록 한다. 최종 분산 행렬은 식 (13)과 같이 계산된다.

$$\Sigma_i = [\phi_{i1}\ \phi_{i2} \cdots \phi_{in}]\, diag(\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{in})\, [\phi_{i1}\ \phi_{i2} \cdots \phi_{in}]^{-1} \tag{13}$$

자세한 내용은 [10]을 참고하면 된다. GK 방법을 FCM에 적용한 클러스터링 결과가 그림 5에 나타나 있다.

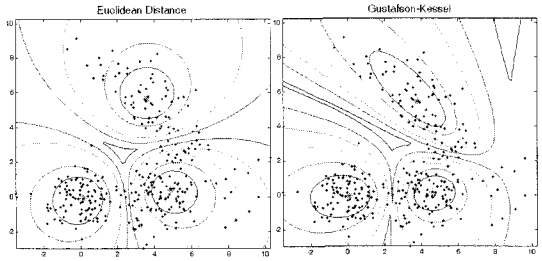

그림 5. 유클리드 거리(왼쪽)와 GK 방법(오른쪽)에 의한 클러스터링
Fig. 5. Clustering result using Euclidean(left) and GK distance(right)

그림 5에 사용된 데이터는 2개의 원형 가우스 분포와 1개의 타원형 가우스 분포에서 생성된 것으로, 기존 FCM의 경우 타원형의 클러스터를 정확히 찾아낼 수 없음을 알 수 있다. 이에 비해 GK 방법은 타원형을 클러스터도 정확히 찾아내고 있다.

3.2 Gath-Geva 방법

GK 방법이 타원형의 클러스터도 찾아낼 수 있지만,

기본적으로 GK 방법은 모든 클러스터들에 속하는 데이터 포인트의 개수가 동일한 것으로 가정한다. 이를 보완하기 위해서는 클러스터에 속하는 포인트의 개수를 고려하여야 하며 이는 클러스터의 사전 확률을 통해 반영이 가능하다. Gath-Geva 방법[11]은 클러스터의 사전 확률까지 고려하여 클러스터 중심과 데이터 포인트 사이의 거리를 식 (14)와 같이 정의한다.

$$d_{ik}^2 = \frac{[\det(\Sigma_i)]^{1/2}}{p_i} \exp\left((x_k - c_i)^T \Sigma_i^{-1} (x_k - c_i)\right) \qquad (14)$$

이 때 Σ_i는 각 클러스터의 퍼지 분산 행렬로 식 (10)으로 계산되고, 식 (12), (13)을 이용하여 특이점이 발생하지 않는 분산 행렬을 구하여 사용하였다. p_i는 각 클러스터의 사전 확률로 식 (15)과 같이 정의된다.

$$p_i = \frac{1}{N} \sum_{k=1}^{N} u_{ik} \qquad (15)$$

그림 6은 두 개의 가우스 분포에서 각각 500개와 50개의 포인트를 생성하여 GK 방법과 GG 방법을 FCM에 적용하여 클러스터링한 결과를 보여주고 있다. 그림에서 알 수 있듯이 GK 방법은 클러스터에 속하는 포인트의 개수에 차이가 많은 경우, 클러스터의 중심이 밀집된 지역으로 치우쳐서 나타나는 경향이 있음을 알 수 있다. 이에 비해 GK 방법은 사전 확률을 고려하므로 두 개의 클러스터가 정확하게 분리되고 있음을 알 수 있다. 또 한 가지 다른 점은, GK 방법은 클러스터의 형태에 중점을 두는 반면 GG 방법은 클러스터들 사이의 경계에 중점을 둔다는 점이다.

이러한 GG의 방법의 특징은 분류 목적으로 사용하기 위해서는 유용하지만 클러스터의 형태를 묘사하기 위해서는 부적합하다. 하지만 PFCM에 GG 방법을 적용하는 경우 계산되는 소속도 값은 분류 목적으로 사용할 수 있고, 전형성은 클러스터의 형태를 알아내기 위해 사용할 수 있다. 다음 장에서는 PFCM에 GG 방법을 사용한 결과를 보이고 이를 GK 방법 및 유클리드 거리를 사용한 PFCM 방법과 비교한다.

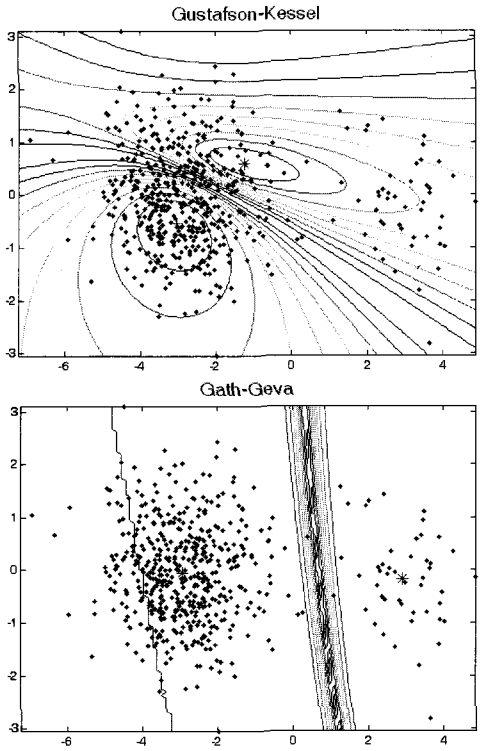

그림 6. 서로 다른 밀도를 갖는 클러스터의 클러스터링 결과 (위 : GK, 아래 : GG)
Fig. 6. Clustering result of two Gaussian clusters with different densities (top : GK, bottom : GG)

Ⅳ. 실험 및 결과 분석

GG 방법이 GK 방법이나 유클리드 거리를 사용하는 방법에 비해 나은 결과를 보이는 것을 확인하기 위해 이 장에서는 3가지 알고리즘(PFCM, PFCM-GK, PFCM-GG)에 2가지 테스트 데이터 집합을 적용하여 실험하였다. 첫 번째 데이터는 그림 5에서와 같이 3개의 가우스 분포에서 생성된 것으로 각 컴포넌트는 100개씩의 포인트를 생성한다(Set I). 두 번째 데이터는 그림 6에서와 같이 2개의 가우스 분포에서 생성된 것으로 각각 500개와 50개의 포인트를 생성한다(Set II).

그림 7, 8, 9는 Set I에 대해 3개의 알고리즘을 수행한 결과를 나타낸다. PFCM의 경우 타원형의 클러스터는 찾아내지 못하는 문제점이 있고, PFCM-GK의 경우 전형

성이 클러스터 중심에서 멀어질 때 급격히 감소하는 것을 알 수 있다. 이에 비해 PFCM-GG는 가우스 분포와 유사한 형태의 전형성을 가지며 소속도는 클러스터들 사이의 경계를 강조하고 있음을 알 수 있다. PFCM-GK와 PFCM-GG는 사용하고자 하는 목적에 따라 선택될 수 있으며 Set I의 경우 거의 동일한 성능을 보여주고 있다.

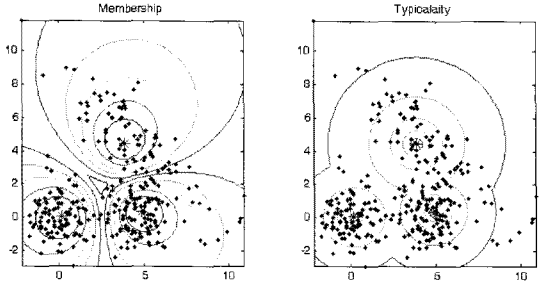

그림 7. PFCM으로 Set I을 클러스터링한 결과
Fig. 7. Clustering result of Set I using PFCM

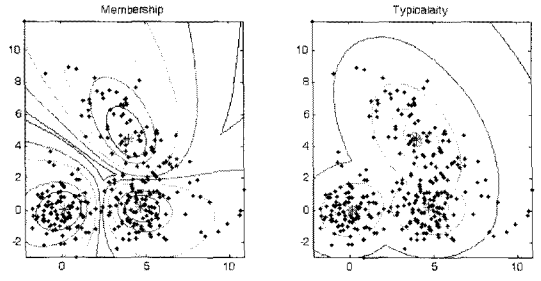

그림 8. PFCM-GK로 Set I을 클러스터링한 결과
Fig. 8. Clustering result of Set I using PFCM-GK

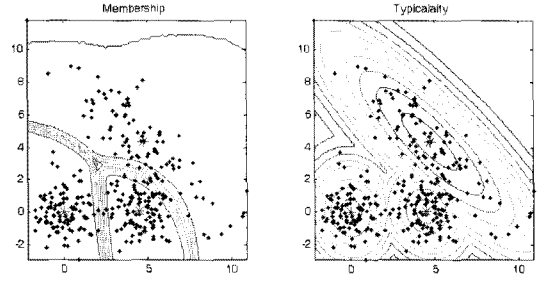

그림 9. PFCM-GG로 Set I을 클러스터링한 결과
Fig. 9. Clustering result of Set I using PFCM-GG

그림 10, 11, 12는 Set II에 3개의 알고리즘을 수행한 결과는 보여주고 있다. Set I과 달리 Set II는 각 클러스터에 속하는 데이터 포인트의 수에 큰 차이가 나므로 클러스터의 밀도를 고려한 PFCM-GG 만이 중심을 바르게 찾아내고 있다. PFCM과 PFCM-GK는 밀도가 높은 클러스터 쪽으로 클러스터의 중심이 치우쳐 발생하고 있다. PFCM-GG의 전형성 값에서 오른쪽 클러스터가 수평으로 늘어난 것은 두 개의 클러스터 크기가 동일한데 비해 밀도는 확연히 차이가 남으로 인한 현상이다.

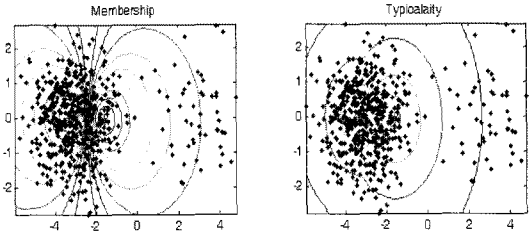

그림 10. PFCM으로 Set II를 클러스터링 한 결과
Fig. 10. Clustering result of Set II using PFCM

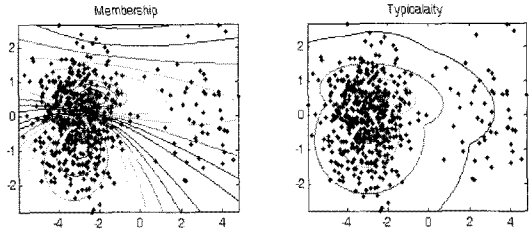

그림 11. PFCM-GK로 Set II를 클러스터링 한 결과
Fig. 11. Clustering result of Set II using PFCM-GK

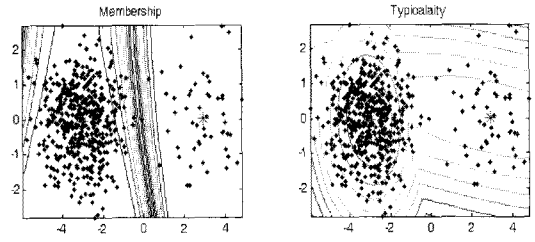

그림 12. PFCM-GG로 Set II를 클러스터링 한 결과
Fig. 12. Clustering result of Set II using PFCM-GG

그림 13은 동일한 데이터에 대해 EM 알고리즘을 사용하여 가우스 혼합 모형을 찾아낸 결과와 PFCM-GG의 전형성를 나타낸 것이다. 그림에서 두 분포는 거의 동일함을 알 수 있다. 이는 그림 8에서 PFCM-GK의 전형성과 비교해보면 명확히 알 수 있다.

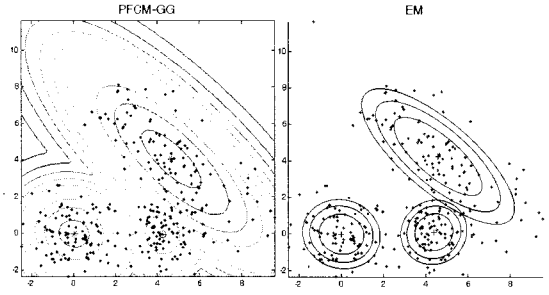

그림 13. PFCM-GG의 전형성과 EM 알고리즘을 이용해 추정한 가우스 분표의 비교
Fig. 13. Comparison of the typicality distribution of PFCM-GG and the Gaussian distribution from EM

표 1은 실제 각 클러스터의 실제 중심과 PFCM-GG와 EM 알고리즘이 찾아낸 클러스터의 중심을 비교하여 나타낸 것으로 두 알고리즘이 찾아낸 중심은 거의 동일함을 알 수 있다. 엔트로피를 목적 함수에 사용하는 퍼지 클러스터링이 가우스 혼합 모형과 연관성이 있음은 잘 알려져 있지만[15], 전형성과의 연관성은 아직 알려진 바가 없으며, 현재 이러한 연관성에 대해서 연구 중에 있다.

표 1. 각 알고리즘이 추정한 클러스터 중심
Table 1. Estimated center positions from each algorithm

Original		PFCM-GG		EM	
X	Y	X	Y	X	Y
0.0000	0.0000	0.0678	-0.0758	0.0832	-0.0735
4.5000	0.0000	4.6467	-0.0556	4.6848	0.0424
4.5000	4.5000	4.9386	4.2967	4.9320	4.2537

V. 결론

퍼지 클러스터링은 클러스터링을 위해 널리 사용되는 방법 중 하나로, 이 논문에서는 소속도와 전형성을 동시에 계산하는 PFCM 알고리즘을 GG 방법을 사용하여 확장하였다. 제안한 방법은 PFCM의 장점과 GG 방법의 장점을 결합하여, 분류 목적에 적합한 소속도와 확률 밀도 추정에 적합한 전형성을 보여준다. GG 방법은 GK 방법과 마찬가지로 유클리드 거리를 사용하는 경우 찾아낼 수 없는 타원형 형태의 클러스터를 찾아낼 수 있으며, 더불어 GK 방법으로는 해결할 수 없는 밀도가 다른 클러스터들도 다룰 수 있는 장점이 있다.

이러한 장점에도 불구하고 제안한 방법은 여전히 개선의 여지가 있다. 실제 데이터는 간단한 테스트 데이터와 달리 잡음과 특이점(outlier)이 많이 존재한다. 비록 PFCM 알고리즘이 전형성을 통해 이들의 영향을 줄일 수 있지만 잡음이 많은 상황에 대처하기 위해서는 잡음 클러스터(Noise Cluster)의 도입이나 로버스트 통계(robust statistics)의 사용 등을 고려할 필요가 있다. 또 다른 개선 방향으로는 임의의 형태를 갖는 클러스터들을 찾아내기 위한 방법으로의 확장이다. 현재 PFCM 및 그 변형들은 가우스 형태의 클러스터들만을 다룰 수 있다. 현재 이러한 목적으로 커널(kernel)을 이용하는 방법이 연구되고 있으며 이러한 커널 기반의 방법은 PFCM에도 적용될 수 있을 것이다. 현재 이러한 개선안 및 PFCM-GG의 전형성과 가우스 분포 사이의 관계에 대해 연구 중에 있다.

참고문헌

[1] L. A. Zadeh, "Fuzzy sets," Information and Control Vol.8, No.3, pp. 338-353, 1965

[2] E. H. Ruspini, "A new approach to clustering," Information and Control 16, pp. 22-32, 1969

[3] Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, Vol.16, No.3, pp. 645-678, 2005

[4] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters," Journal of Cybernetics 3, pp. 32-57, 1974

[5] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, 1981

[6] R. Krishnapuram and J. M. Keller, "A Possibilistic Approach to Clustering," IEEE Transactions on Fuzzy Systems Vol.1, No.2, pp. 98-110, 1993

[7] R. Krishnapuram and J. M. Keller, "The Possibilistic C-Means Algorithm: Insights and Recommendations," IEEE Transactions on Fuzzy Systems Vol.4, No.3, pp. 385-393, 1996

[8] N. R. Pal, K. Pal, J. M. Keller and J. C. Bezdek, "A Possibilistic Fuzzy c-Means Clustering Algorithm," IEEE Transactions on Fuzzy Systems Vol.13, No.4, pp. 517-530, 2005

[9] D. E. Gustafson and W. C. Keller, "Fuzzy clustering with a fuzzy covariance matrix," Proceedings of the 1978 IEEE Conference on Decision and Control, pp. 761-766, 1979

[10] R. Babuska, P. J. van der Veen and U. Kaymak, "Improved Covariance Estimation for Gustafson-Kessel Clustering," Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, pp. 1081-1085, 2002

[11] I. Gath and A.B. Geva, "Unsupervised Fuzzy Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence Vol.11, No.7, pp. 773-781, 1989

[12] B. Ojeda-Magana, R. Ruelas, M. A. Corona-Nakamura and D. Andina, "An Improvement to the Possibilistic Fuzzy C-Means Clustering Algorithm," World Automation Congress 2006 (WAC '06), pp. 1-8, 2006

[13] N. R. Pal, K. Pal and J. C. Bezdek, "A mixed c-means clustering model," Proceedings of IEEE International Conference on Fuzzy Systems, pp. 11-21, 1997

[14] J. S. Zhand and Y. W. Leung, "Improved Possibilistic C-Means Clustering Algorithms," IEEE Transactions on Fuzzy Systems Vol.12, No.2, pp. 209-217, 2004

[15] D. Tran and M. Wagner, "Fuzzy Entropy Clustering," Proceedings of the 9th IEEE International Conference on Fuzzy Systems, pp.152-157, 2000

저자소개

허경용(Gyeongyong Heo)

1994년 2월:연세대학교 전자공학과 (공학사)
1996년 8월:연세대학교 본대학원 전자공학과(공학석사)
2004년 9월~현재 : Dept. of Computer and Information Science and Engineering, University of Florida
※ 관심분야 : Machine Learning, Bayesian Network, Image Processing,

최세운 (Sewoon Choe)

2001년 2월: 홍익대학교 전자전기공학과 (공학사)
2004년 5월: Dept. of Elec. and Comp. Eng., Univ of Florida (공학석사)
2006년 1월 ~ 현재: Dept. of Biomedical Engineering, University of Florida
※ 관심분야: Medical Image Processing, Non-Invasive Cancer Imaging, Monte Carlo Simulation

우영운(Young Woon Woo)

1989년 2월 : 연세대학교 전자공학과(공학사)
1991년 8월 : 연세대학교 본대학원 전자공학과(공학석사)
1997년 8월 : 연세대학교 본대 학원 전자공학과 (공학박사)
1997년 9월 ~ 현재: 동의대학교 멀티미디어공학과 교수
2008년 ~ 현재: 한국해양정보통신학회 학술이사
※ 관심분야: 지능시스템, 패턴인식, 퍼지이론