Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

Possibilistic Fuzzy C-Means 클러스터링 알고리즘의 확장

Extension of the Possibilistic Fuzzy C-Means Clustering Algorithm

허경용[1], 우영운[2], 김광백[3]

[1] Computer and Information Science and Engineering, University of Florida, US
E-mail: hgycap@hotmail.com
[2] 부산시 부산진구 동의대학교 멀티미디어공학과
E-mail: ywwoo@deu.ac.kr
[3] 부산시 사상구 신라대학교 컴퓨터정보공학부
E-mail: gbkim@silla.ac.kr

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 각 포인트가 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 확장한다. 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 있다는 사실을 실험 결과를 통해 확인할 수 있었다.

Key Words : Clustering, Fuzzy Clustering, PFCM, Gustafson-Kessel Method, Gath-Geva Method

1. 서 론

Fuzzy C-Means(FCM)[1]는 퍼지 클러스터링 방법 중 널리 사용되는 방법이지만 하나의 포인트에 대한 소속도의 합이 1이 되도록 함으로써 때때로 직관적인 분할과는 다른 결과를 보여주며 잡음에 민감한 특성을 갖는다. 이러한 문제를 해결하기 위해 Krishnapuram과 Keller[2]는 FCM의 제약을 제거한 Possibilistic C-Means(PCM)를 제안하였다. 하지만 PCM은 각각의 클러스터들이 서로 영향을 주지 않으므로 초기화에 민감하고 때때로 중복된 클러스터를 찾아내는 문제점이 있다[3]. 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[4, 5] 그 중 하나가 Possibilistic Fuzzy C-Means(PFCM)[4]이다.

이들 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)[6]에 의해 제안되었다. 또 한 가지 널리 사용되는 방법은 Gath와 Geva (GG)[7]에 의해 제안된 방법으로 가우스 분포 함수에 반비례하는 값을 거리로 사용한다.

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. GG 방법을 사용한 결과는 기존의 PFCM 방법[4], PFCM에 GK를 사용한 방법[8]과의 비교를 통해 그 차이점을 명확히 알 수 있었다.

2. 클러스터링

2.1 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 (1)과 같으며, 식에서 두 번째 항은 모든 t_{ik} 값이 0이 되는 경우 목적 함수가 최소화되는 자명해(trivial solution)를 제거하기 위해 첨가된 항이다.

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

이 때 δ_i는 각 클러스터의 부피를 나타내는 값으로, 클러스터의 크기 추정과 특이점(outlier) 판별에 영향을 미친다. PCM은 FCM과 마찬가지로 AO(alternating optimization) 방법을 통해 국부 최대값을 구하며, 전형성 t_{ik}는 식 (2)와 같이 주어진다[2].

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

식 (2)는 한 데이터 포인트와 하나의 클러스터 중심 사이의 거리만을 고려한다. 즉 각 클러스터들에 소속되는 정도는 독립적으로 결정되며, 이처럼 클러스터들 사이의 상관관계를 고려하지 않음으로 해서 중첩된 클러스터 문제를 유발한다.

그림 1은 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.

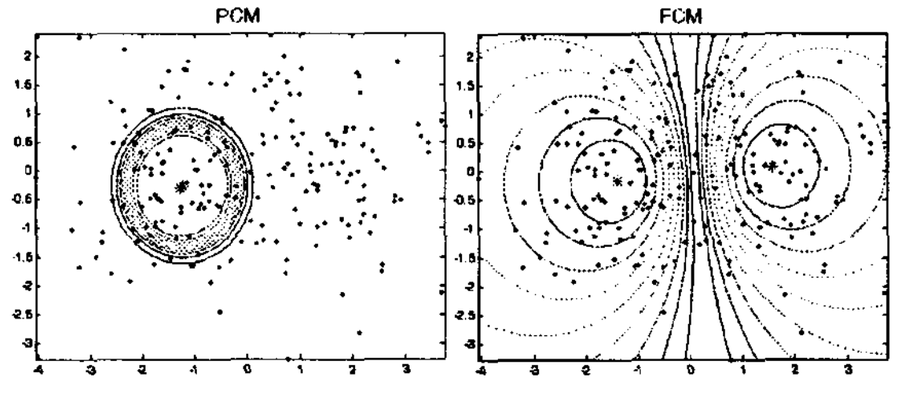

그림 1. 중첩된 클러스터에 대한 클러스터링 결과
(왼쪽 : PCM, 오른쪽 : FCM)

2.2 Possibilistic Fuzzy C-Means (PFCM)

FCM과 PCM은 각각의 장점과 단점이 있으므로, 서로의 단점을 보완하기 위한 방법으로 소속도와 전형성을 함께 사용하고자 하는 시도가 있어왔고[4, 5] 그 중 하나가 PFCM[4]이다. PFCM의 목적 함수는 식 (3)과 같이 주어진다.

$$J_{PFCM} = \sum_{i=1}^{C}\sum_{k=1}^{N}(au_{ik}^{m} + bt_{ik}^{\eta})d_{ik}^{2} + \sum_{i=1}^{c}\delta_i \sum_{k=1}^{N}(1-t_{ik})^{\eta} \quad (3)$$

이 때 a, b는 소속도와 전형성에 대한 가중치 상수이며, η는 소속도에서 m과 동일한 역할을 전형성에서 하는 상수이다.

그림 2는 그림 1과 동일한 데이터를 PFCM을 통해 클러스터링한 결과이다. PFCM은 클러스터 중심의 계산을 위해 소속도와 전형성을 동시에 고려함으로써 FCM의 잡음 민감성을 완화하고 PCM의 중첩된 클러스터 문제를 해결한다.

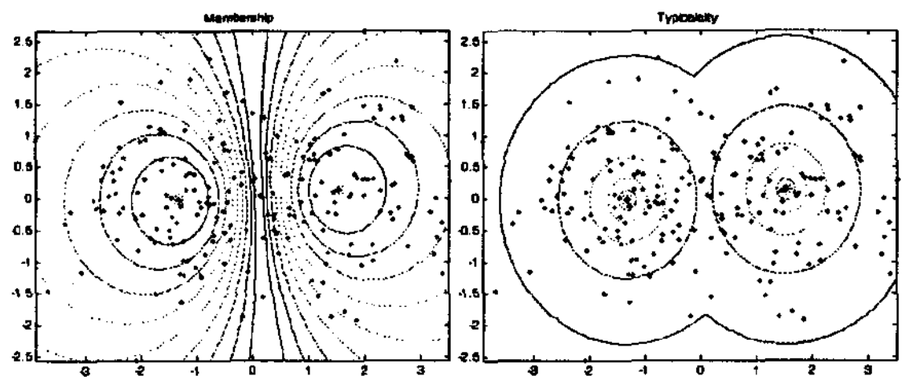

그림 2. PFCM을 이용한 클러스터링 결과
(왼쪽 : 소속도, 오른쪽 : 전형성)

3. 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)[6]에 의해 제안되었

다. 다른 한 가지 방법은 Gath와 Geva (GG)[7]에 의해 제안된 방법으로 가우스 분포 함수에 반비례하는 값을 거리로 사용한다.

3.1. Gustafson-Kessel 방법

Gustafson-Kessel 방법은 식 (4)로 정의되는 퍼지 분산 행렬 Σ_i를 통해 각 클러스터가 타원형의 분포를 가질 수 있도록 해준다.

$$\Sigma_i = \frac{\sum_{k=1}^{N} u_{ik}^m (x_k - c_i)(x_k - c_i)^T}{\sum_{k=1}^{N} u_{ik}^m} \quad (4)$$

대한 자세한 내용은 [6]을 참고하면 된다. GK 방법을 FCM에 적용한 클러스터링 결과는 그림 3과 같다.

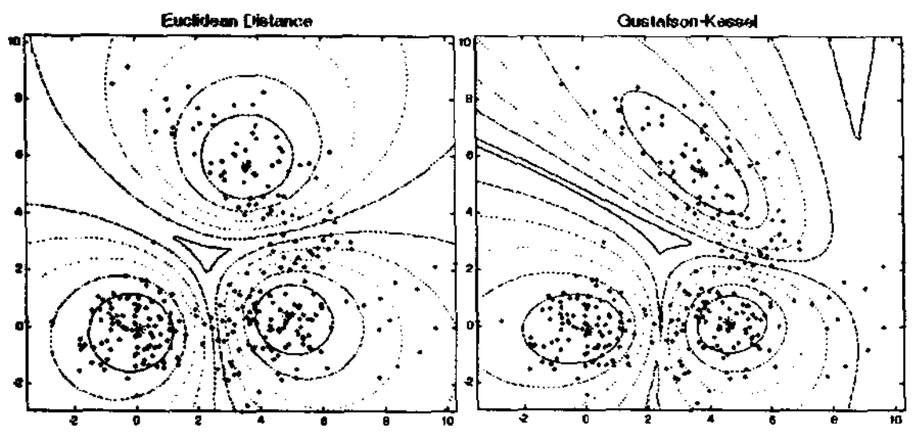

그림 3. 유클리드 거리(왼쪽)와 GK 방법(오른쪽)에 의한 클러스터링

그림 3에 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

Gath-Geva 방법[7]은 클러스터의 사전 확률까지 고려하여 클러스터 중심과 데이터 포인트 사이의 거리를 식 (5)와 같이 정의한다.

$$d_{ik}^2 = \frac{[\det(\Sigma_i)]^{1/2}}{p_i} \exp\left((x_k - c_i)^T \Sigma_i^{-1} (x_k - c_i)\right) \quad (5)$$

이 때 Σ_i는 각 클러스터의 퍼지 분산 행렬로 식 (4)로 계산된다. p_i는 각 클러스터의 사전 확률로 식 (6)과 같이 정의된다.

$$p_i = \frac{1}{N} \sum_{k=1}^{N} u_{ik} \quad (6)$$

그림 4는 두 개의 가우스 분포에서 각각 500개와 50개의 포인트를 생성하여 GK 방법과 GG 방법을 FCM에 적용하여 클러스터링한 결과를 보여주고 있다. 그림에서 알 수 있듯이 GG 방법은 GK 방법과 달리 클러스터의 경계를 강조하고 있음을 알 수 있다. 또한 클러스터에 속하는 포인트의 개수에 차이가 많은 경우, GK 방법은 클러스터의 중심이 데이터 포인트가 밀집된 지역으로 치우쳐서 나타나는 경향이 있는 반면, GK 방법은 사전 확률을 고려하므로 두 개의 클러스터가 정확하게 분리되고 있음을 알 수 있다.

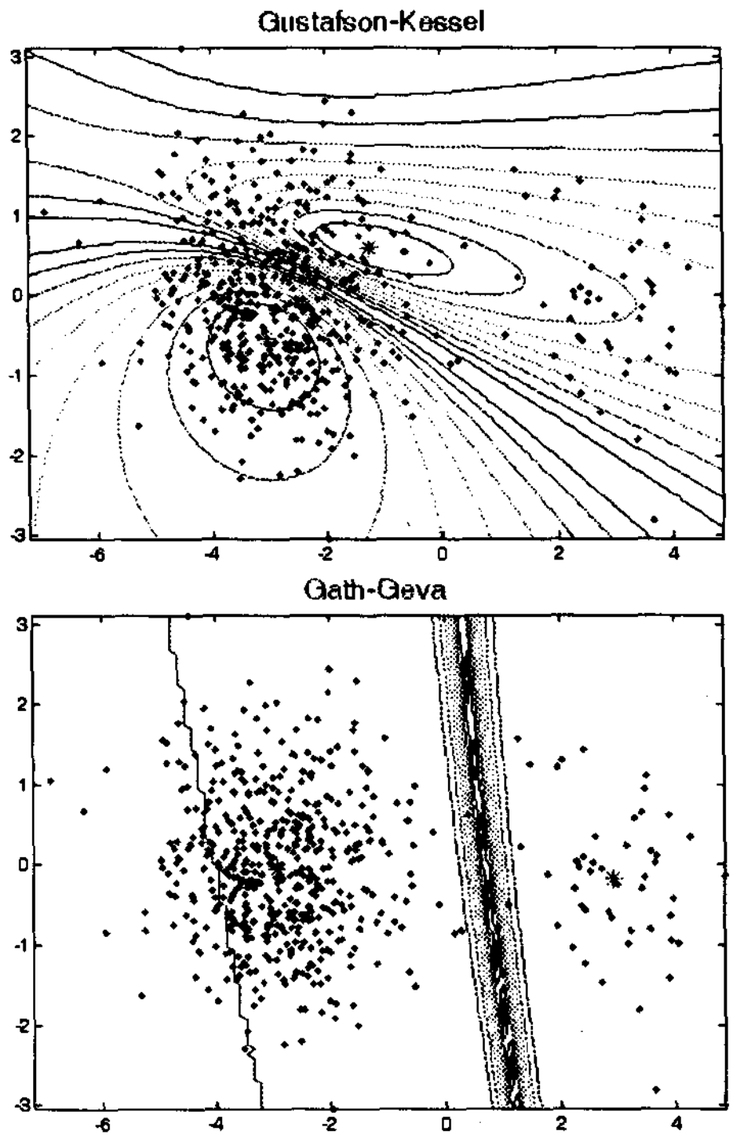

그림 4. 서로 다른 밀도를 갖는 클러스터의 클러스터링 결과 (위 : GK, 아래 : GG)

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.

4. 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 테스트 데

이터 집합을 적용하여 실험하였다. 테스트 데이터는 그림 4에서와 같이 2개의 가우스 분포에서 생성된 것으로 각각 500개와 50개의 포인트를 가진다.

그림 5, 6, 7은 테스트 데이터에 3개의 알고리즘을 수행한 결과는 보여주고 있다. 그림에서 알 수 있듯이 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 나기 때문에 발생하는 현상이다.

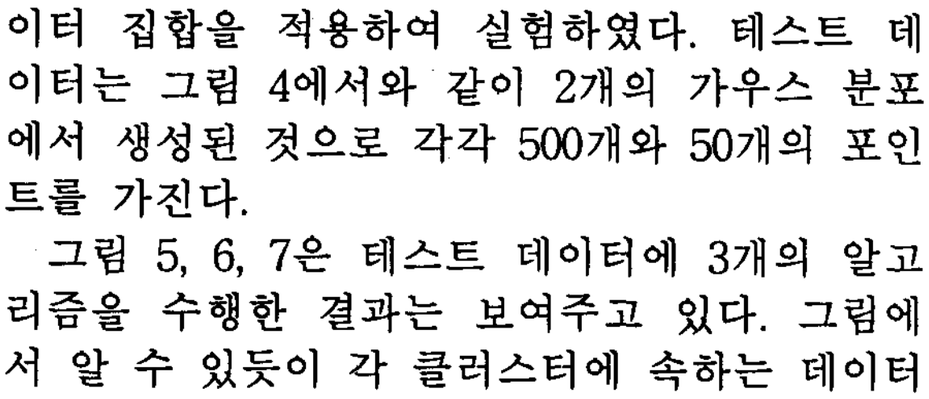

그림 5. PFCM으로 클러스터링한 결과

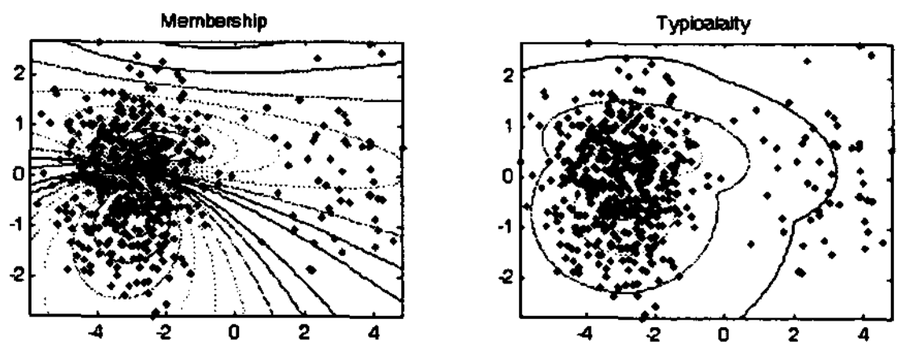

그림 6. PFCM-GK로 클러스터링한 결과

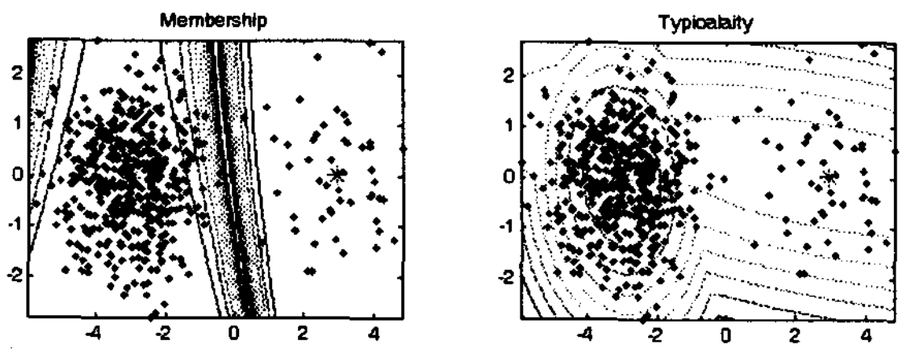

그림 7. PFCM-GG로 클러스터링한 결과

5. 결 론

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 장점을 확인할 수 있었다.

참 고 문 헌

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.