Proceedings of KFIS Fall Conference, 2003. 5. 3

개선된 Gustafson-Kessel 알고리즘을 이용한 퍼지 클러스터링

Fuzzy Clustering with Improving Gustafson-Kessel Algorithm

김승석, 곽근창, 유정웅, 전명근
충북대학교 전기전자공학부

Sung-Suk Kim, Keun-Chang Kwak, Jeong-Woong Ryu, Myung-Geun Chun
School of Electrical and Electronics.
Chungbuk National University
E-mail : powerkimss@hotmail.com

요 약

본 논문에서는 Gaussian Mixture Model을 이용한 Gustafson-Kessel 알고리즘의 성능을 개선하였다. 분포 및 밀도가 다른 데이터에 대하여 적절한 클러스터 파라미터를 추정함으로써 클러스터링의 성능을 개선한다. 일반적인 클러스터링 알고리즘의 경우, 데이터가 편중되거나 각 데이터의 밀도가 서로 틀린 경우 클러스터의 파라미터가 정확하게 클러스터를 표현하지 못하는 문제점을 가지고 있다. 제안된 방법에서는 Gustafson-Kessel 알고리즘을 이용하여 클러스터 파라미터를 추정하며 알고리즘내의 파라미터 일부를 Gaussian Mixture Model을 이용하여 동적으로 갱신하였다. 시뮬레이션을 통하여 제안된 방법의 유용성을 보인다.

1. 서론

클러스터링 알고리즘은 적은 클러스터 파라미터로 큰 데이터의 특성 등을 표현할 수 있어 다양한 분야에서 활발하게 연구되고 있으며 또한 적용되고 있다[1][2]. 일반적인 퍼지 클러스터링 알고리즘의 경우 정상적인 데이터 분포를 갖는 경우 클러스터 파라미터 추정에 좋은 성능을 가지는 반면 편중된 데이터나 밀도가 서로 다른 데이터의 클러스터 파라미터가 데이터를 제대로 표현하지 못하는 문제점을 가진다[3][4]. 클러스터 파라미터 추정을 위한 방법으로 유클리언 거리를 이용하거나 분산 등을 추가하여 반복적인 수행을 통하여 클러스터를 추정한다. 제안된 방법은 클러스터 중심을 추정하기 위하여 유클리디언 거리를 이용하며 또한 공분산을 행렬을 응용한 Mahalanobis 거리를 이용한 Gustafson-Kessel (G-K) 알고리즘을 이용하였다[3][5]. 또한 G-K 알고리즘 내에서의 체적을 표현하는 ρ의 값을 추정하기 위하여 Gaussian Mixture Model (GMM)의 알고리즘을 추가하여 G-K 알고리즘의 성능을 개선하였다[6-8].

제안된 방법의 유용성을 예제 시뮬레이션을 통하여 보였다.

2. G-K 알고리즘과 GMM

2.1 Gustafson-Kessel알고리즘

Fuzzy C-Means (FCM)이 가지는 문제점에 대하여 성능을 확장한 것이 G-K 알고리즘이다[3-5]. 이 알고리즘은 데이터 집합에서 서로 다른 기하학적인 형태의 클러스터를 검출하기 위하여 유클리디언 거리측정과 함께 적응적인 거리 측정을 적용한 표준 FCM의 확장 형태이다. 각각의

클러스터는 다음과 같은 내적 노름을 가지는 유도된 행렬 A_i를 가진다.

$$D^2_{ikA_i}=(z_k-v_i)^T A_i(z_k-v_i) \quad (1)$$

여기서 행렬 A_i는 c-mean 함수에서 최적화 변수로 이용되며, 데이터의 위상학적 구조에서 각 클러스터에 대하여 거리 측정에 적응적이다. A를 유도된 행렬 $A=(A_1, A_2, \ldots, A_c)$ c개의 조합으로 기술한다. 여기서 GK 알고리즘의 목적함수는 다음과 같이 정의한다.

$$J(Z; U, V, A)=\sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m D^2_{ikA_i} \quad (2)$$

여기서, $U\in M_{fc}$, $V\in \mathbf{R}^{n\times c}$, $m>1$ 이고, 이는 다음과 같이 해를 구할 수 있다.

$$(U, V, A)=\arg\min_{M_{fc}\times \mathbf{R}^{n\times c}\times PD^n} J(Z; U, V, A) \quad (3)$$

여기서 PD^n은 $n\times n$의 양의 크기로 정의된 행렬의 공간이다. 고정된 A에 대하여, 다음 식은 직접적으로 적용된다.

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{c}(D_{ikA}/D_{jkA})^{2/(m-1)}},\ 1\le i\le c,\ 1\le k\le N \quad (4)$$

$$v_i=\frac{\sum_{k=1}^{N}(\mu_{ik})^m z_k}{\sum_{k=1}^{N}(\mu_{ik})^m};\ 1\le i\le c \quad (5)$$

여기서 목적함수는 A_i에 대하여 직접적으로 최소화 할 수 없다. 가능한 해를 구하기 위해서, A_i는 몇가지 방법들을 이용하여 한정시킨다. 이것을 수행하기 위한 일반적인 방법은 A_i의 행렬식을 한정하는 것이다.

$$|A_i|=\rho_i,\ \rho_i>0,\ \forall i \quad (6)$$

라그랑지 곱의 방법을 이용하면, A_i에 대하여 다음과 같은 표현을 얻는다.

$$A_i=[\rho_i det(F_i)]^{-1/n} F_i^{-1} \quad (7)$$

여기서 F_i는 i번째 클러스터의 Fuzzy 공분산 행렬로 다음과 같이 정의된다.

$$F_i=\frac{\sum_{k=1}^{N}(\mu_{ik})^m (z_k-v_i)(z_k-v_i)^T}{\sum_{k=1}^{N}(\mu_{ik})^m} \quad (8)$$

여기서 공분산은 U에서 소속도 정도에 의해 가중 처리된다. GK 알고리즘은 다음과 같다.

Gustafson-Kessel 알고리즘

주어진 데이터 집합 Z에 대하여, 클러스터의 수 $1\le c\le N$와 가중 멱지수 $m>1$과 종료 조건 $\varepsilon>0$을 정한다.

$U^{(0)}\in M_{fc}$와 같이 초기 분할 행렬을 임의로 생성한다.

Step 1 : 중심을 계산한다.

$$v_i^{(l)}=\frac{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m z_k}{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m},\ 1\le c\le N$$

Step 2 : 클러스터 분할 행렬을 계산한다.

$$F_i=\frac{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m (z_k-v_i^{(l)})(z_k-v_i^{(l)})^T}{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m},\ 1\le$$

Step 3 : 거리를 계산한다.

$$D^2_{ikA_i}=(z_k-v_i^{(l)})^T\times [(\rho_i det(F_i)^{1/n} F_i^{-1}](z_k-v_i^{(l)})$$

Step 4 : 분할 행렬을 갱신한다.

만약 $1\le i\le c$, $1\le k\le N$에 대하여 $D_{ikA_i}>0$ 이면

$$\mu_{ik}^{(l)}=\frac{1}{\sum_{j=1}^{c}(D_{ikA_i}/D_{jkA_j})^{2/(m-1)}}$$ 이고 아니면,

$D_{ikA_i}>0$, $\sum_{i=1}^{c}\mu_{ik}^{(l)}=1$ 에서 $\mu_{ik}^{(l)}\in[0,1]$ 이면 $\mu_{ik}^{(l)}=0$ 이다.

이 과정을 $\|U^{(l)}-U^{(l-1)}\|<\varepsilon$ 이 만족할 때까지 반복한다.

2.3 Gaussian Mixture Model

G-K 알고리즘에서는 능동적으로 ρ를 결정하지 않으므로 제안된 방법에서는 이를 결정하기 위한 방법으로 GMM을 이용하였다[6][7]. 일반적인 GMM에서 확률밀도함수는 다음 식에 의해 주어진다.

$$p(x|\Phi) = \sum_{j=1}^{c} p(x|w_j, \Phi)P(w_j) \quad (9)$$

여기서 w_j는 성분들이고, $P(w_j)$는 그것들의 사전 확률 (prior probability) 혹은 혼합 파라미터들이다 [9]. 그리고 $p(x|w_j, \Phi)$는 파라미터 Φ의 벡터가 알려졌을 때 성분의 확률 밀도이다. 샘플 $X=\{x_k\}$의 log 우도(likelihood)는 다음과 같다.

$$L(\Phi|X) = \sum_k \log p(x_k|\Phi)$$
$$= \sum_k \log \sum_j p(x_k|w_j, \Phi)P(w_j) \quad (10)$$

이 식은 직접적인 해를 가지지 못한다. 이러한 문제는 EM 알고리즘을 사용함으로서 해결할 수 있다. EM 알고리즘은 두 단계로 수행되는데, 간략히 그 과정을 살펴보면, 먼저 E-단계에서 사후확률 (posterior probability)을 다음과 같이 베이즈의 정리 (Bayes's Theorem)를 이용하여 계산한다.

$$P(w_j|x_k, \Phi) = \frac{p(x_k|w_j, \Phi)P(w_j)}{\sum_l p(x_k|w_l, \Phi)P(w_l)} \equiv h_{kj} \quad (11)$$

만약 성분밀도가 d-차원 가우시안이 되도록 선택하면, $p(x|w_j, \Phi) \sim N_d(\mu_j, \Sigma_j)$은 다음과 같다.

$$p(x|w_j, \Phi) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \times \exp\left[-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right] \quad (12)$$

그러면 식(3)은 식(4)로 표현될 수 있으며 이는 가우시안 확률을 이용하는 식(5)로 표현되어질 수 있다.

$$h_{kj} = \frac{g_j|\Sigma_j|^{-1/2}\exp\left[-\frac{1}{2}(x_k-\mu_j)^T\Sigma_j^{-1}(x_k-\mu_j)\right]}{\sum_l g_l|\Sigma_l|^{-1/2}\exp\left[-\frac{1}{2}(x_k-\mu_l)^T\Sigma_j^{-1}(x_k-\mu_l)\right]} \quad (13)$$

단, $g_j \equiv P(w_j)$ 이다.

다음에 M-단계에서 성분 파라미터 Φ의 성분을 식 (14)~(15)과 같이 갱신한다.

$$\mu_j^{t+1} = \frac{\sum_k h_{kj}x_k}{\sum_k h_{kj}} \quad (14)$$

$$\Sigma_j^{t+1} = \frac{\sum_k h_{kj}(x_k-\mu_j^{t+1})(x_k-\mu_j^{t+1})^T}{\sum_k h_{kj}} \quad (15)$$

$$g_j^{t+1} = \frac{1}{n}\sum_k h_{kj} \quad (16)$$

여기서 μ_j, Σ_j, g_j 각각 EM 알고리즘에 의하여 갱신되는 가우시안 확률의 평균과 분산, 가중치로 표현할 수 있으며 가중치를 ρ의 값으로 이용하였다.

3. 시뮬레이션 및 결과

시뮬레이션 데이터로는 일반적인 실험 데이터가 아닌 임의로 생성된 데이터를 이용하여 성능을 평가하였다. 생성된 데이터는 크기 및 밀도 형태를 변형하였으며 각각의 데이터를 조합하여 통하여 전체 데이터를 만들었다.

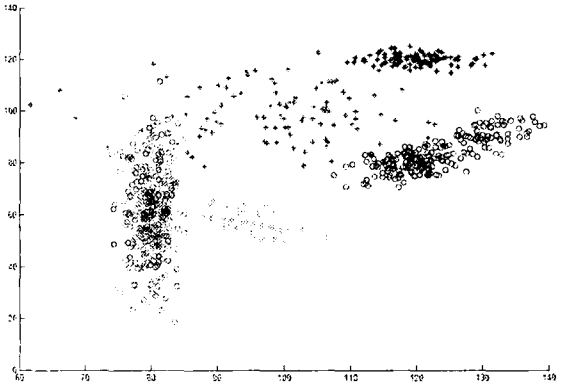

그림 1. 데이터 분포

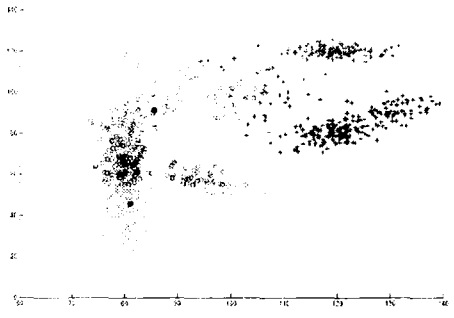

그림 2. FCM의 추정결과

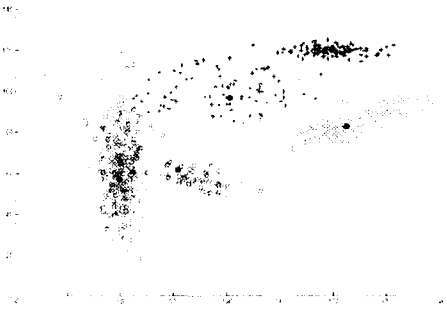

그림 3. G-K 알고리즘의 추정 결과

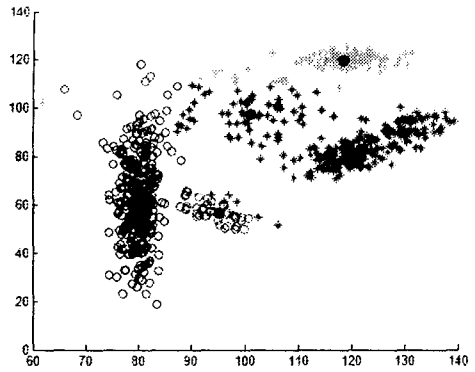

그림 4. 제안된 방법에 의한 추정 결과

표 1. 성능 평가

	FCM	G-K	제안된 방법	비고
1번째	42	36	**39**	
2번째	153	17	**0**	
3번째	0	0	**0**	
4번째	0	0	**0**	
5번째	39	50	**12**	
종합	234	103	**51**	

전체 데이터의 크기는 750개이며 각각의 클러스터는 데이터의 분포와 밀도가 서로 다르다. 그림 1에서 보면 각 클러스터는 데이터의 분포와 밀도가 서로 다르며 그 형태 역시 상이하다. 그림 2에서 볼 수 있듯이 일반적인 FCM의 경우 단지 유클리디언 거리만을 이용함으로써 목적함수에 의하여 클러스터의 밀도가 높은 방향으로 파라미터가 치중하게 된다. 그림 3에서, G-K 클러스터링은 각 클러스터의 파라미터가 같은 체적으로 가지려는 방향으로 클러스터를 추정한다. 그림 4에서, 클러스터의 유클리디언 거리와 공분산 행렬 및 체적까지 고려한 방법으로 위의 알고리즘 결과 보대 좋은 성능을 보였다. 표 1에서도 알 수 있듯이 제안된 방법의 클러스터 결과가 FCM 및 G-K 알고리즘보다 좋은 성능을 보임을 알 수 있다.

4. 결론

본 논문에서는 G-K 알고리즘의 클러스터링 성능 개선을 위한 방법을 시도하였다. 일반적인 FCM의 특성을 확장한 G-K 알고리즘에서 체적을 표현하는 ρ를 제안된 방법에 의하여 추정함으로써 시뮬레이션에서의 성능 개선을 보았다. 주어진 데이터는 각 클러스터의 수와 분포가 서로 다르면서도 인접하여 일반적인 클러스터 추정 방법에 대하여 문제점을 가지고 있었다. 제안된 방법은 클러스터의 할당 및 체적에 관한 문제점에 대하여 GMM을 이용한 파라미터 추정을 G-K 알고리즘에 넣어 좀더 좋은 성능을 볼 수 있었다. 향후 연구과제로는 적응적인 ρ 결정 문제의 개선 및 클러스터 밀도 및 분포 등의 다양한 조건을 동시에 만족하는 클러스터 추정 방법 제안 및 뉴로-퍼지 모델로의 적용을 통하여 지속적인 성능 개선 등이 있다.

6. 참고문헌

[1] J. S. R. Jang, C. T. Sun, E. Mizutani, Neuro-Fuzzy and Soft Computing : A Computational Approach to Learning and Machine Intelligence, Prentice Hall, 1997.

[2] Timothy J. Ross, Fuzzy Logic with Engineering Application, McGraw-Hill, 1995

[3] Robert Babuska, Fuzzy Modeling for Control, Kluwer Academic, 1998

[4] Uzay Kaymak, Magne Setnes, "Fuzzy Clustering With Volume Prototypes and Adaptive Cluster Merging", IEEE Trans on Fuzzy Systems, Vol.10, No. 6, pp.705-712, 2002

[5] Raghu Krishnapuram, Jongwoo Kim, "A Note on the Gustafson-Kessel and Adaptive Fuzzy Clustering Algorithms", IEEE Trans on Fuzzy Systems, Vol. 7, No. 4, 1999

[6] Todd. K. Moon, "The Expectation-Maximization Algorithm", IEEE Signal Processing Magazine, 1996

[7] 김승석, 곽근창, 유정웅, 전명근, "GMM과 클러스터링 기법에 의한 뉴로-퍼지 시스템 모델링", 한국퍼지및지능시스템학회, Vol. 12, No. 6, pp.571-576, 2002

[8] Guorong Xuan, Wei Zhang, Peiqi Chai, "EM algorithm of Gaussian Mixture Model and Hidden Harkov Model", Image Processing, Proceedings, International Conference on, Vol. 1, pp. 145-148, 2001.

[9] Simon Haykin, Neural Network : A Comprehansive Foundation, Prentice Hall, 1999