커널 기반의 Possibilistic C-Means 클러스터링 알고리즘

A Kernel based Possibilistic C-Means Clustering Algorithm

최길수, 최병인, 이정훈
한양대학교 전자공학과

Kil-Soo Choi, Byung-In Choi, Frank Chung-Hoon Rhee
Department of Electronic Engineering, Hanyang University, Ansan, Korea
E-mail : {kschoi, bichoi, frhee}@fuzzy.hanyang.ac.kr

요 약

Fuzzy Kernel C-Means(FKCM) 알고리즘은 커널 함수를 통하여 구형의 데이터뿐만 아니라 Fuzzy C-Means(FCM)에서는 분류하기 힘든 복잡한 형태의 분포를 갖는 데이터를 분류할 수 있다. 하지만 FCM과 같이 노이즈에 대해서는 민감한 성질을 가진다. 이처럼 노이즈(noise)에 민감한 성질을 보완하기 위해서 본 논문에서는 Possibilistic C-Means 알고리즘에 커널 함수를 적용하였다. 본 논문에서 제안된 Kernel Possibilistic C-Means(KPCM) 알고리즘은 일반적인 데이터에 대해 FKCM과 같은 성능의 클러스터링 수행이 가능하며 노이즈가 있는 데이터에 대해서는 FKCM보다 더욱 정확한 클러스터링을 수행할 수 있다.

Key Words : Kernel, FKCM, KPCM, Noise

1. 서론

일반적으로 잘 알려진 Fuzzy C-Means(FCM)은 유클리디안 공간상에서의 거리를 이용하여 퍼지 소속도를 할당해 줌으로써 클러스터링을 수행하게 된다[5]. 대부분의 구형 데이터에 대해서 FCM은 좋은 성능의 클러스터링을 수행하지만 유클리디안 거리를 기준으로 하므로 고리형 데이터와 같이 복잡한 형태의 분포를 가지는 데이터들에 대해서는 클러스터링이 불가능하다. 이러한 단점을 극복하기 위해서 Fuzzy Kernel C-Means(FKCM)가 제안되었다[2]. 이것은 FCM에 커널 함수를 적용하여 데이터를 입력 공간이 아닌 커널 속성(feature) 공간으로 변환하여 클러스터링을 수행하는 것을 목적으로 한다. 커널 기반의 알고리즘, 즉 FKCM은 구형의 데이터뿐만 아니라 복잡한 형태의 분포를 갖는 데이터에 대해서도 정확한 클러스터링이 가능하다.

그러나 FKCM은 FCM과 같이 패턴과 각 클러스터 센터 사이의 거리에 대한 퍼지 소속도의 합이 1이어야 한다. 그러므로 노이즈와 같은 패턴에 대해서도 다른 패턴과 같이 퍼지 소속도를 할당하기 때문에 경계면의 위치가 영향을 받게 된다는 것을 알 수 있다. 이러한 단점을 보완하기 위해서 본 논문에서는 Kernel Possibilistic C-Means(KPCM)를 제안하였다. KPCM은 노이즈 패턴에 대해서는 다른 패턴과 비교하여 상대적으로 작은 퍼지 소속도를 할당하기 때문에 노이즈가 포함된 데이터에 대해서 더 뛰어난 클러스터링 수행이 가능할 수 있다.

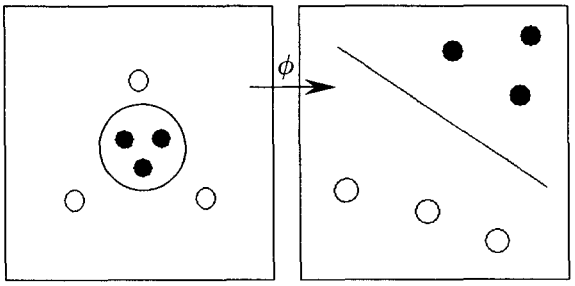
본 논문은 다음과 같이 구성된다. 두 번째 절에

접수일자 : 2004년 1월 1일
완료일자 : 2012년 12월 31일
감사의 글 : 본 연구는 한국과학기술원 영상정보특화연구센터를 통한 국방과학연구소의 연구비 지원으로 수행되었습니다.

서 커널 함수에 대해 설명하고 세 번째 절에서는 FKCM에 대하여 소개하고 네 번째 절에서 본 논문에서 제안하고 있는 KPCM에 대해 설명한다. 다음으로 FCM과 FKCM, KPCM의 결과를 비교하고 마지막으로는 논문에 대한 결론을 언급하도록 한다.

2. 커널 함수

커널의 기본 목적은 공간 변환 함수를 사용하여 입력 데이터들을 입력 속성 공간을 커널 함수를 통한 커널 속성 공간으로 변환하여 주는 것이다.



(a) Input feature space (b) Kernel feature space

그림 1. 커널 함수에 의한 공간 변환

그림1과 같이 데이터들의 공간을 변환하기 위하여 공간 변환 함수를 사용한다. 즉 입력 공간에서의 데이터를 $\mathbf{X}_i$, i=1,…,N 이라 한다면 함수를 통해 커널 속성 공간으로 변환된 데이터는 $\phi(\mathbf{X}_j)$, j=1,…,N 로 나타낼 수 있다. 여기에서 $\phi()$는 입력 공간의 데이터를 비선형적으로 커널 속성 공간으로 변환시켜주는 함수이다. 이렇게 정의된 변환 함수에 의해 두 함수값 간의 내적(inner product)을 커널 함수로서 정의하고[1], 다항식이나 가우시안 등의 함수를 사용할 수 있다.

$$K(X, Y) = \phi(X) \cdot \phi(Y) = (X \cdot Y + b)^d \quad (1)$$

$$K(X, Y) = e^{-\frac{(X-Y)^2}{2\sigma^2}} \quad (2)$$

(1)과 (2)에서 d는 다항식의 차수, b는 상수, σ^2은 클러스터 분산(variance)을 나타내는 파라미터이다. 커널 함수를 사용함으로써, 두 벡터에 대한 변환 함수값을 구하지 않고 커널 함수의 값을 직접 구할 수 있다. 입력 공간상에서 $\mathbf{X}_i$와 $\mathbf{X}_j$의 커널 속성 공간상 거리는 커널 함수에 의해 (3)과 같이 표현된다.

$$\begin{aligned} d_{ij}^2 &= \|\phi(X_i) - \phi(X_j)\|^2 \\ &= \phi(X_i)\phi(X_i) - 2\phi(X_i)\phi(X_j) \\ &\quad + \phi(X_j)\phi(X_j) \\ &= K(X_i, X_i) - 2K(X_i, X_j) + K(X_j, X_j) \end{aligned} \quad (3)$$

3. Fuzzy Kernel C-Means(FKCM)

일반적인 FCM의 경우와 마찬가지로 FKCM의 경우에도, 다음의 목적 함수를 최소화하는 것을 기초로 한다.

$$J(X; U, V) = \sum_{j=1}^{C}\sum_{i=1}^{N} u_{ij}^m d^2(X_i, V_j), \quad 2 \le C < N \quad (4)$$

여기에서 $\mathbf{U}$는 각 클러스터에 대한 소속도, u_{ij}를 원소로 가지는 행렬을 나타내고 m은 퍼지화의 정도를 나타내는 상수로서 $m \in (1,\infty)$인 조건을 만족하는 상수이다. 또한 $\mathbf{U}$는 다음의 조건을 만족해야 한다.

$$\sum_{j=1}^{C} u_{ij} = 1, \ \forall i \text{ and } 0 < \sum_{i=1}^{N} u_{ij} < N, \ \forall j$$

(4)식의 목적함수를 최소화하는 소속도를 구해보면 다음과 같은 식으로 표현된다.

$$u_{ij} = \frac{(1/d^2(X_i, V_j))^{1/(m-1)}}{\sum_{j=1}^{C}(1/d^2(X_i, V_j))^{1/(m-1)}}, \quad j=1,\cdots,C, \ i=1,\cdots,N \quad (5)$$

패턴과 센터의 거리, $d^2(\mathbf{X}_i,\mathbf{V}_j)$은 커널 함수를 이용하여 다음과 같이 나타낼 수 있다.

$$d^2(X_i, V_j) = K(X_i, X_i) - 2K(X_i, V_j) + K(V_j, V_j) \quad (6)$$

초기 센터에 대해 모든 데이터와 센터간의 초기 소속도가 식 (5)에 의해서 결정되면 패턴과 센터 사이의 새로운 거리는 다음 식들을 이용하여 갱신할 수 있다.

$$K(X_i, \hat{V}_j) = \frac{\sum_{k=1}^{N}(u_{jk})^m K(X_k, X_i)}{\sum_{k=1}^{N}(u_{jk})^m} \quad (7)$$

$$K(\hat{V}_j, \hat{V}_j) = \frac{\sum_{k=1}^{N}\sum_{l=1}^{N}(u_{jk})^m (u_{jl})^m K(X_k, X_l)}{\left(\sum_{k=1}^{N}(u_{jk})^m\right)^2} \quad (8)$$

$$\text{where } \phi(\hat{V}_j) = \frac{\sum_{i=1}^{N}(u_{ji})^m \phi(X_i)}{\sum_{i=1}^{N}(u_{ji})^m} \quad (9)$$

이러한 거리를 이용하여 (5)식의 목적 함수를 최소화 시켜주는 소속도를 구할 수 있다.

이와 같은 과정을 초기에 정의한 종료조건을 만족할 때까지 반복하여 FKCM을 수행한다.

4. Kernel Possibilistic C-Means(KPCM)

KPCM의 기본적인 목적은 서론에서 간단히 설명했듯이 커널 함수를 이용한 공간 변환을 통하여 클러스터링을 수행하던 FKCM보다 노이즈에 대해 강한 성질을 부여하는 것이다. 그래서 다음과 같이 FCM보다 노이즈에 강한 Possibilistic C-Means(PCM)의 목적 함수식을 이용한다.

$$J(X;U,V)=\sum_{j=1}^{C}\sum_{i=1}^{N}u_{ij}^{m}d^2(X_i,V_j)+\sum_{j=1}^{C}\eta_j\sum_{i=1}^{N}(1-u_{ij})^m \quad (10)$$

위의 식에서 η_j는 (12)식에 의해 구해지는 적절한 양수이다. η_j의 값에 의해 KPCM에서는 노이즈 데이터에 대해서 FKCM보다 상대적으로 작은 소속도를 할당함으로써 노이즈 데이터에 강한 성능을 나타낸다. PCM에서는 적절한 초기 센터를 구하기 위해서 먼저 FCM을 수행한다. 마찬가지로 KPCM에서도 FKCM을 먼저 수행한다.

주어진 목적 함수가 최소값을 갖도록 해주는 소속도는 다음과 같이 표현할 수 있다.

$$u_{ij}=\frac{1}{1+\left(\frac{d^2(X_i,V_j)}{\eta_j}\right)^{1/(m-1)}} \quad (11)$$

여기에서 η_j는 다음의 식에 의해서 구할 수 있다.

$$\eta_j=\frac{\sum_{i=1}^{N}u_{ij}^{m}d^2(X_i,V_j)}{\sum_{i=1}^{N}u_{ij}^{m}} \quad (12)$$

이렇게 정의된 소속도에 대하여 (6)식과 (7), (8), (9)식에 의해 커널 속성 공간상에서의 패턴과 센터의 거리 $d^2(\mathbf{X}_i,\mathbf{V}_j)$를 갱신한다. 위의 과정을 정의된 종료조건을 만족할 때까지 반복하면 주어진 목적함수를 최소화 할 수 있는 각 패턴들의 소속도를 구할 수 있다. 본 논문에서 제시하고 있는 KPCM(Kernel Possibilistic C-Means) 알고리즘을 정리하면 다음과 같다.

Kernel Possibilistic C-Means Clustering

Step 1. Initialization of the membership

(Using FKCM Algorithm)

Initialize random k center $\mathbf{V}_k$ and m , σ^2;

Compute the initial membership $u_{ij}^{(0)}$;

Step 2. Minimization of the objective function

(Kernel Possibilistic C-Means Algorithm)

Compute $d^2(X_i,\hat{V}_j)$ using (6),(7),(8),(9);

Do : Calculate η_{ij} using (12);

Compute $u_{ij}^{(l)}$ using (11);

Update $d^2(X_i,\hat{V}_j)$ using (6),(7),(8),(9);

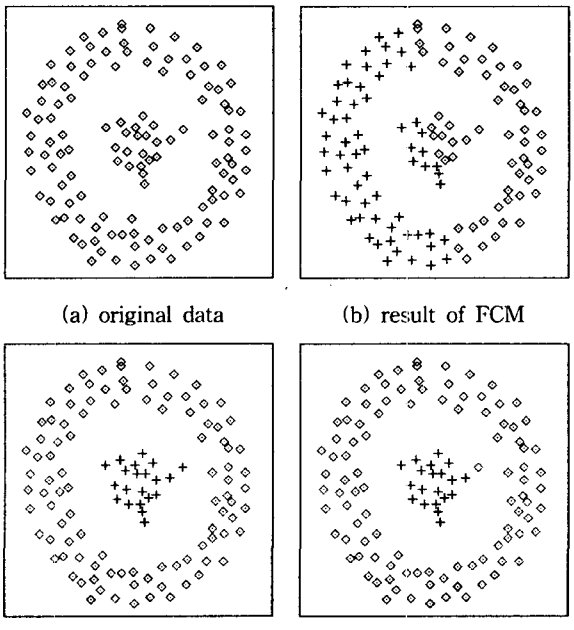
Until : $u_{ij}^{(l+1)}-u_{ij}^{(l)}<\epsilon$;

5. 실험 결과

본 절에서는 고리 모양과 T 모양의 데이터에 대해 FCM과 (2)식의 가우시안 커널 함수를 사용하는 FKCM과 KPCM의 결과를 비교하도록 한다. FKCM과 KPCM에서 필요한 파라미터 중 퍼지화 정도를 나타내는 m값은 1.5를 사용하였고 σ^2은 각 데이터에 따라 임의의 값을 사용하였다.

5.1 Ring Data

고리 모양의 데이터는 각각 두 개의 속성들과 클러스터들을 가지는 120개의 패턴들로 구성된다. FKCM의 σ^2은 0.007을 사용하였고 KPCM은 0.06의 값을 사용하였다. 그림 2에서 첫 번째 실험의 결과를 살펴보면 고리 모양의 데이터는 FCM으로는 클러스터링이 불가능하다. 하지만 FKCM과 KPCM의 경우에는 내부와 외부의 두 클러스터가 잘 분류된다는 것을 알 수 있다.



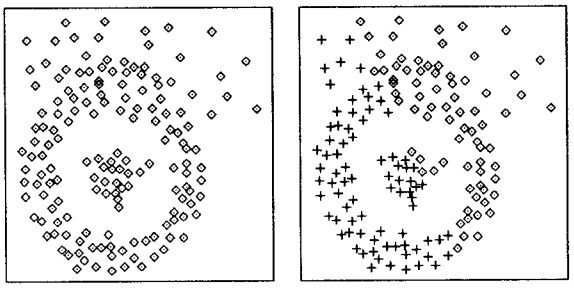
(a) original data (b) result of FCM

(c) result of FKCM (d) result of KPCM

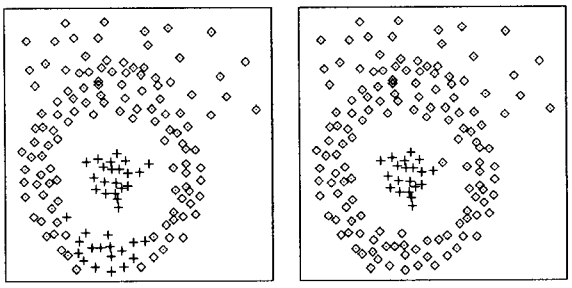
그림 2. 노이즈가 없는 고리 모양의 데이터

5.2 Ring Data with Noises

두 번째 실험은 첫 번째 실험의 고리모양 데이터에 36개의 노이즈가 포함된 데이터에 대해서 같은 과정을 수행한다. 각 파라미터는 첫 번째 실험과 동일하게 설정하였다. 그림 3의 결과에서 보듯이 FKCM의 경우, 노이즈가 포함되었을 때에는 그림3.(c)와 같이 내부의 클러스터가 노이즈의 영향을 받는다는 것을 알 수 있다. 하지만 KPCM은 그림3.(d)와 같이 노이즈가 없을 때와 같은 클러스터링 결과를 가진다.



(a) noisy data (b) result of FCM

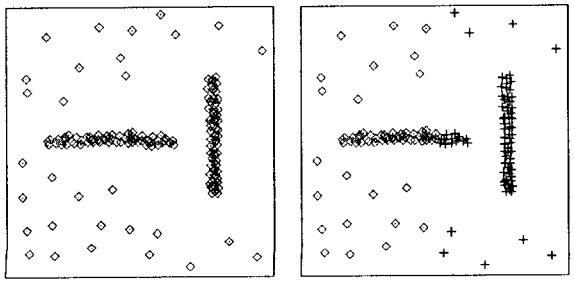


(c) result of FKCM (d) result of KPCM

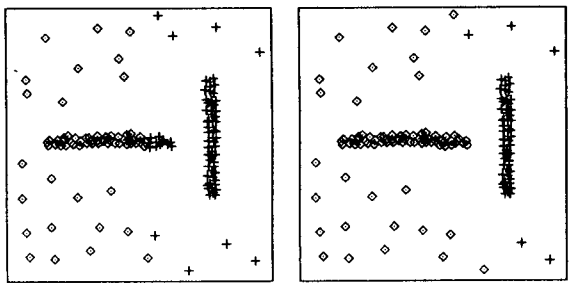
그림 3. 노이즈가 있는 고리 모양의 데이터

5.3 T-Shape Data with Noises

세 번째 실험은 노이즈가 있는 T형(구형)의 데이터에 대해서 결과를 비교한다. 실험에 쓰인 데이터는 110개의 패턴에 대해 30개의 노이즈 패턴을 가진 데이터이다. FKCM과 KPCM의 σ^2은 모두 1.0을 사용하였다. 그림 4의 결과에서 보듯이 T형(구형) 데이터의 경우 FCM과 FKCM은 두개의 클러스터를 정확하게 분류하지 못하는 것을 알 수 있다. 하지만 KPCM의 경우 그림4.(d)에서와 같이 노이즈에 상관없이 두 개의 클러스터를 잘 구분할 수 있다는 것을 알 수 있다.



(a) noisy data (b) result of FCM



(c) result of FKCM (d) result of KPCM

그림 4. 노이즈가 있는 T 모양의 데이터

6. 결론

본 논문에서는 커널 함수를 사용하여 KPCM 알고리즘을 제안하였다. FCM에 커널 함수를 적용한 FKCM은 일반적인 구형 모양의 데이터 뿐만 아니라 고리 모양의 데이터에 대해서도 클러스터링이 가능하였다. 하지만 FCM과 마찬가지로 노이즈에 대해서는 민감한 현상을 나타내었다. 본 논문에서 제안된 KPCM은 실험 결과에서 보듯이 복잡한 형태의 분포를 가지는 데이터에 대한 클러스터링이 가능하고 FKCM에서보다 노이즈에 대해 좋은 클러스터링 결과를 가진다는 것을 알 수 있다.

7. 참고문헌

[1] M. Girolami, "Mercer Kernel-Based Clustering in Feature Space," *IEEE Trans. Neural Networks.* vol. 13, no. 5, pp. 780-784, May 2002.

[2] Z. Wu, W. Xie, J. Yu, "Fuzzy C-Means Clustering Algorithm based on Kernel Method," *IEEE Conf. Computational Intelligence and Multimedia Applications*, pp. 49-54, September 2003.

[3] R. Krishnapuram, J. Keller, "A Possibilistic Approach to Clustering," *IEEE Trans. Fuzzy syst.* vol. 1, no. 2, pp. 98-110, May 1993.

[4] R. Krishnapuram, J. Keller, "A Possibilistic C-Means Algorithm : Insights and Recommendations," *IEEE Trans. Fuzzy syst.* vol. 1, no. 2, pp. 98-110, May 1993.

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

[6] M. Barni, V. Cappellini, A. Mecocci, "Comments on "A Possibilistic Approach to Clustering" ," *IEEE Trans. Fuzzy Syst.* vol. 4, no. 3, August 1996.