계층적인 구조를 이루는 KPCM 알고리즘

A Hierarchy of Kernel PCM-Generated Clusters

구양협, 최병인, 이정훈
한양대학교 전자전기제어계측공학부

Yang-Hyup Koo, Byung-In Choi and Frank Chung-Hoon Rhee
School of Electrical Engineering Computer Science, Hanyang University
E-mail : {yhkoo, bichoi, frhee}@fuzzy.hanyang.ac.kr

ABSTRACT

커널함수를 이용한 클러스터링 방법은 일반적인 목적함수 기반의 클러스터링 방법에 비해 고리모양과 같은 복잡한 모양의 데이터를 클러스터링할 때 훨씬 효율적이다. 그러나, 커널기반의 클러스터링 방법은 거리함수를 계산하기 위하여 커널함수를 연산해야 하기 때문에 클러스터 수가 많아지면, 일반적인 목적함수 기반의 클러스터링 방법에 비하여 계산량이 급격히 증가하는 단점이 있다. 따라서, 본 논문에서는 이러한 단점을 개선하기 위하여 커널기반의 클러스터링 기법에 계층적인 클러스터링 모델을 적용한다.

Key words : 커널, 클러스터링, 계층 구조, KPCM, FKCM

I. 서 론

클러스터링 알고리즘으로 일반적으로 사용되는 Fuzzy C-Means(FCM) 기법은 구형의 데이터에 대해서 좋은 성능의 클러스터링을 수행하지만, 고리 모양과 같은 복잡한 데이터에 대해서는 클러스터링 수행이 어렵다[4]. 또한, 잡음이 섞여있는 데이터를 클러스터링할 경우 결과가 좋지 않을 수 있다. 이러한 이유는 유클리디언 공간상에서 패턴과 클러스터 센터간의 거리함수에 따라 클러스터들간의 상대적인 소속도를 할당하기 때문이다[5]. 이러한 단점들을 극복하기 위하여, Kernel Possibilistic C-Means(KPCM)이 제안되었다[1]. 데이터의 입력 속성공간을 커널 속성공간으로 변환하여 클러스터링을 수행하는 것을 목적으로 하기 때문에 구형의 데이터뿐만 아니라 복잡한 형태를 갖는 데이터에 대해서도 적절한 클러스터링이 가능하다. 그러나, KPCM 기법은 패턴과 클러스터 센터간의 거리함수를 커널 속성 공간에서 연산해 주어야 하므로, 일반적인 목적함수 기반의 클러스터링 방법에 비하여 클러스터 수가 많아짐에 따라 계산량이 급격히 증가하는 단점이 있다. 따라서 본 논문에서는 기존의 KPCM 기법에 클러스터링을 몇 단계로 나누어 계층적으로 클러스터링을 수행하는 방법을 적용할 것이다. 이러한 계층적인 클러스터링 방법은 클러스터 수가 많아질 때 계산량이 급격히 증가하는 현상을 완화할 수 있다[2]. 또한, 클러스터링을 여러 단계로 나누어 수행하기 때문에 커널함수에서 사용되는 분산 파라미터를 각 단계의 데이터 분포에 맞게 적절히 할당할 수 있다. 따라서, 기존의 KPCM 보다 뛰어난 성능을 가질 수 있다.

본 논문은 다음과 같은 순서로 구성된다. 먼저, 두 번째 절에서는 커널 함수, KPCM 및 본 논문에서 제안하는 Hierarchical Kernel Possibilistic C-Means(HKPCM)에 대해 설명하고, 세 번째 절에서는 실험을 통하여 본 논문에서 제안하는 방법과 다른 클러스터링 알

고리즘들간의 성능을 비교 분석할 것이며 마지막으로 네 번째 절에서 결론을 맺을 것이다.

II. 본 론

2.1 커널 함수

커널의 기본 목적은 공간 변환 함수를 사용하여 입력 데이터들의 입력 속성 공간을 커널 함수를 통한 커널 속성 공간으로 변환하여 주는 것이다. 입력 공간에서의 데이터를 $\mathbf{x}_j$, $j=1,\cdots,n$ 이라 한다면 함수를 통해 커널 속성 공간으로 변환된 데이터는 $\phi(\mathbf{x}_j)$, $j=1,\cdots,n$ 로 나타낼 수 있다. 이렇게 정의된 변환 함수에 의해 두 함수값 간의 내적(inner product)을 커널 함수로서 정의하고, 다항식이나 가우시안 등의 함수를 사용할 수 있다[3].

$$K(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})\cdot\phi(\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}+b)^d \tag{1}$$

$$K(\mathbf{x},\mathbf{y}) = e^{\frac{(\mathbf{x}-\mathbf{y})^2}{2\sigma^2}} \tag{2}$$

(1)과 (2)에서 d는 다항식의 차수, b는 상수, σ^2은 분산 파라미터를 나타낸다. 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.

$$\begin{aligned} d_{ij}^2 &= \left\|\phi(\mathbf{x}_i)-\phi(\mathbf{x}_j)\right\|^2 \\ &= \phi(\mathbf{x}_i)\phi(\mathbf{x}_i) - 2\phi(\mathbf{x}_i)\phi(\mathbf{x}_j) + \phi(\mathbf{x}_j)\phi(\mathbf{x}_j) \\ &= K(\mathbf{x}_i,\mathbf{x}_i) - 2K(\mathbf{x}_i,\mathbf{x}_j) + K(\mathbf{x}_j,\mathbf{x}_j) \end{aligned} \tag{3}$$

2.2. Kernel Possibilistic C-Means(KPCM)

PCM과 마찬가지로 KPCM도 (4)식의 목적함수를 최소화하는 방향으로 클러스터링이 수행된다[1].

$$\begin{aligned} J(\mathbf{x};U,V) = &\sum_{j=1}^{C}\sum_{i=1}^{N} u_{ij}^m d^2(\mathbf{x}_i,\mathbf{v}_j) \\ &+ \sum_{j=1}^{C}\eta_j\sum_{i=1}^{N}(1-u_{ij})^m \end{aligned} \tag{4}$$

$i = 1,\cdots,N,\ j = 1,\cdots,C$

N = Number of input data and C = Number of clusters

KPCM을 먼저 수행하기 전에 FKCM을 먼저 수행하여 적절한 초기 센터값 및 파라미터 η_j 값을 구한다. PCM과 마찬가지로 목적함수가 최소값을 갖도록 해주는 소속도는 다음과 같이 표현할 수 있다.

$$u_{ij} = \frac{1}{1+\left(\dfrac{d^2(\mathbf{x}_i,\mathbf{v}_j)}{\eta_j}\right)^{1/(m-1)}} \tag{5}$$

여기에서 η_j는 다음의 식에 의해서 구할 수 있다.

$$\eta_j = \frac{\sum_{i=1}^{N} u_{ij}^m d^2(\mathbf{x}_i,\mathbf{v}_j)}{\sum_{i=1}^{N} u_{ij}^m} \tag{6}$$

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

$$\begin{aligned} d^2(\mathbf{x}_i,\mathbf{v}_j) = &K(\mathbf{x}_i,\mathbf{x}_i) - 2K(\mathbf{x}_i,\mathbf{v}_j) \\ &+ K(\mathbf{v}_j,\mathbf{v}_j) \end{aligned} \tag{7}$$

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

$$K(\mathbf{x}_i,\hat{\mathbf{v}}_j) = \frac{\sum_{k=1}^{N}(u_{jk})^m K(\mathbf{x}_k,\mathbf{x}_i)}{\sum_{k=1}^{N}(u_{jk})^m} \tag{8}$$

$$K(\hat{\mathbf{v}}_j,\hat{\mathbf{v}}_j) = \frac{\sum_{k=1}^{N}\sum_{l=1}^{N}(u_{jk})^m (u_{jl})^m K(\mathbf{x}_k,\mathbf{x}_l)}{\left(\sum_{k=1}^{N}(u_{jk})^m\right)^2} \tag{9}$$

$$\text{where } \phi(\hat{\mathbf{v}}_j) = \frac{\sum_{i=1}^{N}(u_{ij}^m)^m \phi(\mathbf{x}_i)}{\sum_{i=1}^{N}(u_{ij})^m} \tag{10}$$

이렇게 정의 소속도에 대하여 (7)식과 (8), (9), (10)식에 의해 커널 속성 공간상에서의 패턴과 센터의 거리 $d^2(\mathbf{x}_i,\mathbf{v}_j)$를 갱신한다. 위의 과정을 정의된 종료조건을 만족할 때까지 반복한다.

2.3. Hierarchical Kernel Possibilistic C-Means(HKPCM)

본 논문에서 제안하는 HKPCM알고리즘의 기본적인 구조를 그림 1과 같이 나타내었다. 그림 1에서도 알 수 있듯이, 클러스터링을 여러 단계로 나누어 수행한다. 먼저, 전체 데이터 집합에 대해서, 미리 정의된 각 레벨의 클러스

터 개수 c만큼 클러스터링을 수행한다. 발생된 각 클러스터들의 성능값을 조사하여 클러스터링 결과가 제일 안 좋은 클러스터가 선택된다.

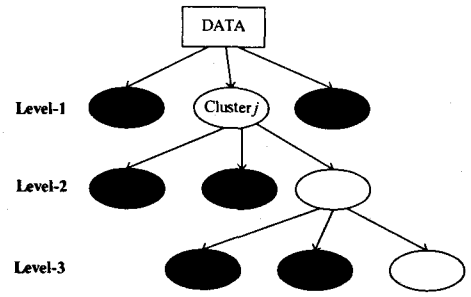

그림 1. HKPCM의 기본 구조

클러스터링에 대한 성능 기준은 여러 가지가 될 수 있다. 본 논문에서는 각 클러스터내의 데이터에 대하여 (4)식의 목적함수를 적용하여 성능값을 구하였다. 따라서, 각 클러스터내의 성능값이 낮을수록 클러스터링이 잘 수행된 것이고, 성능값이 높을수록 클러스터링이 잘못 수행된 것이다. 따라서, 제일 높은 성능값을 가지는 클러스터가 선택된다. 그림 1에서 흰색 원이 각 레벨에서 제일 높은 성능값을 가지는 클러스터이다. 각 레벨에서 클러스터가 선택되면, 그 클러스터에 속하는 데이터를 그 다음 하위 레벨에서 c개의 클러스터로 클러스터링한다. 이러한 방법으로 발생된 클러스터 개수가 미리 정의된 전체 클러스터 개수 C이 될 때까지 하위 레벨로 계속 클러스터링을 수행해 나간다[2]. 물론, 각 레벨에서 클러스터링을 수행할 때, KPCM을 사용하여 클러스터링을 수행한다. 본 논문에서 제시하는 HKPCM 알고리즘을 정리하면 다음과 같다.

HKPCM Clustering Algorithm

Step1) Initialize
Set arbitrarily total number of clusters, C;
Set arbitrarily number of clusters, c in each level ($c < C$);

Step2) Hierarchical Clustering
Perform KPCM for entire data set using c;
Assign number of generated clusters, c' to c;
Do :
Search for cluster having highest performance index;
Perform KPCM for data from selected cluster c;
$c' = c' + c$;
Until : ($c' \neq C$)

KPCM처럼 Fuzzy Kernel C-Means (FKCM)에도 계층적인 클러스터링 방법을 적용하여 수행시간을 단축 시킬 수 있다.

III. 실 험 결 과

본 절에서는 입력 패턴 데이터가 가우시안 분포를 이루고 있는 Gaussian data 와 두개의 고리모양으로 이루어진 Two-ring data를 가지고 실험을 수행한다. 퍼지화를 나타내는 상수 m=1.5로 정의하였고, 각 클러스터링 알고리즘의 종료 조건 값은 ε = 0.00001로 정하였다. 분산 파라미터값은 클러스터링 되어지는 데이터에 따라 적절한 값을 임의로 주었다. 아울러, 우리는 상기한 실험 데이터를 가지고 기존의 커널 기반 클러스터링 방법(KPCM, FKCM)과 본 논문에서 제안하는 계층적인 커널 기반 클러스터링 방법(HKPCM, HKFCM)을 비교분석한다. Hierarchical Kernel Fuzzy C-Means(HKFCM)도 HKPCM과 마찬가지로 FKCM을 계층적으로 클러스터링 한 방법이다.

3.1. Gaussian Data

그림 2의 데이터는 총 8개의 클러스터로 구성 되어져 있으며 각 클러스터 마다 30개씩의 패턴 데이터가 있다.

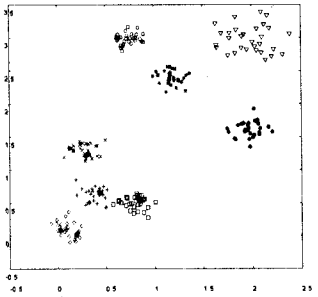

그림 2. Gaussian data

아래 결과를 살펴보면, 알고리즘의 수행 시간을 비교하면, HKPCM의 경우 KPCM보다 14배 정도 수행 시간이 빨라졌으며, HKFCM의 경우 FKCM보다 2배 정도 수행 시간이 빨라졌다는 것을 알 수 있다.

표 1. 각 알고리즘의 성능 비교

	반복회수	수행시간	에러율
FKCM	17	34s	6.3%
HKFCM	522	16s	2.5%
KPCM	284	560s	3.3%
HKPCM	193	39s	5%

3.2. Two-ring Data

그림 3에서 볼 수 있듯이 HKPCM과 HKFCM의 경우 완벽하게 데이터를 클러스터링하는 반면 FKCM과 KPCM의 경우 결과가

좋지 않다는 것을 알 수 있다. FKCM과 KPCM의 경우 분산 파라미터 값을 0.000009에서 1까지 변화시켜가면서 결과를 관찰하였으나 결과가 좋지 않았다. 이러한 이유는 두 개의 링을 나누기 위해 필요한 분산 파라미터 값과 링 내부와 외부를 나누는 분산 파라미터 값이 다르기 때문이다. 반면 HKFCM 및 HKPCM의 경우 각 레벨마다 여러 가지 다른 분산 파라미터값을 적절히 주어 좀 더 정교한 클러스터링이 가능하다. 또한, HKPCM의 경우 수행시간이 72sec가 걸린 반면 PKCM의 경우 평균적으로 195sec 정도 걸렸고, FKCM의 경우 평균적으로 75sec정도 걸린 반면 HKFCM의 경우 19sec정도의 시간밖에 걸리지 않았다.

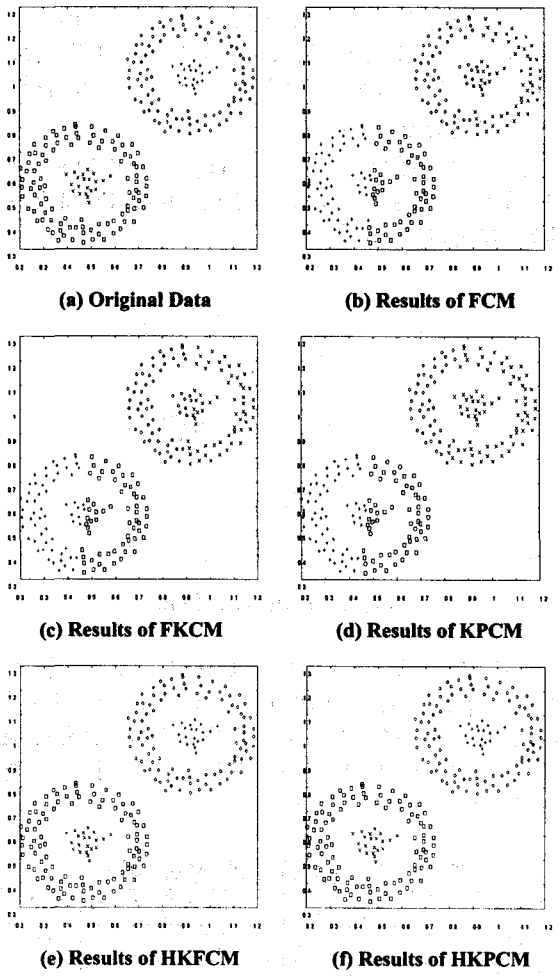

그림 3. Results of two-ring Data

IV. 결 론

기존의 일반적인 목적함수기반의 클러스터링 기법과 달리 커널 기반의 클러스터링 알고리즘은 복잡한 모양의 데이터에 대해서도 클러스터링을 수행할 수 있었다. 그러나 커널 기반의 클러스터링 알고리즘은 클러스터 수가 증가함에 따라 계산량이 급증하는 단점이 있다. 따라서, 본 논문에서는 클러스터링을 몇 단계로 나누어 계층적으로 클러스터링하는 새로운 모델을 제안하였다. 실험 결과에서도 보았듯이, 본 논문에서 제안하는 방법이 기존의 커널 기반의 클러스터링 방법에 비하여 성능이 뛰어나면서 수행속도도 더 빠르다는 것을 알 수 있었다. 그러나, 각 레벨에 적용되는 클러스터들의 최적의 성능 기준 및 커널함수에 적용되는 적절한 분산 파라미터를 선택하는 방법에 대한 연구가 향후 필요하다.

감사의 글: 본 연구는 한국과학기술원 영상정보특화연구센터를 통한 국방과학연구소의 연구비 지원으로 수행되었습니다.

V. 참고문헌

[1] F. Rhee, K. Choi, and B. Choi, "A kernel-based possibilistic C-means clustering algorithm," *in Proc, International Fuzzy Systems Association,* vol. 11, pp. 939-944, July 2005.

[2] A. Pedrycz and M. Reformat, "A hierarchy of FCM-generated clusters," *in Proc, International Fuzzy Systems Association,* vol. 11, pp. 951-956, July 2005.

[3] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Trans. Neural Networks,* Vol. 13, no. 5, pp. 780-784, May 2002.

[4] Z. Wu, W. Xie, and J. Yu, "Fuzzy C-means clustering algorithm based on kernel method," *IEEE Conf. Computational Intelligence and Multimedia Applications,* pp. 49-54, September 2003.

[5] J. Bezdek, *Pattern recognition with fuzzy objective function algorithms,* Plenum Press, New York 1981.