韓國컴퓨터情報學會 論文誌
第16卷 第1號, 2011. 1.

2011-16-1-1-5

Fuzzy c-means의 문제점 및 해결 방안

허 경 용*, 서 진 석**, 이 임 건***

Problems in Fuzzy c-means and Its Possible Solutions

Gyeongyong Heo*, Jinseok Seo**, Imgeun Lee***

요 약

클러스터링은 주어진 데이터 집합을 균일한 특성을 가지는 몇 개의 그룹으로 묶는 대표적인 비교사 학습 방법 중 하나로 지금까지 다양한 형태의 알고리듬이 개발되어 다양한 응용 분야에서 사용되어 왔다. 이 중 fuzzy c-means (FCM)는 분할 기반의 클러스터링 기법에 속하는 알고리듬으로 1970년대에 정립된 이후 지금까지 사용되고 있는 대표적인 클러스터링 알고리듬 중의 하나이다. 하지만 FCM에는 여러 가지 문제점이 있으며 이를 해결하기 위해 지금까지도 다양한 FCM의 변형이 제안되고 있다. 이 논문에서는 먼저 FCM의 문제점을 살펴보고 이를 해결하기 위해 제안된 방법들을 통해 연구 방향을 제시하고자 한다. FCM의 문제점을 해결하고자 하는 대부분의 FCM 변형은 주어진 문제 영역의 지식을 활용하고 있다. 하지만 이 논문에서는 문제 영역을 한정하지 않고 모든 문제에 적용할 수 있는 일반적인 방안을 제시하는데 초점을 둔다. 제시하는 방안은 앞으로 더 많은 연구가 필요하지만 클러스터링을 연구하고자 하는 이들에게 최근의 연구 동향과 더불어 출발점을 제시할 수 있을 것으로 기대한다.

▸ Keyword : Fuzzy c-means, 비교사 학습, 분할 기반 클러스터링, FCM의 문제점

Abstract

Clustering is one of the well-known unsupervised learning methods, in which a data set is grouped into some number of homogeneous clusters. There are numerous clustering algorithms available and they have been used in various applications. Fuzzy c-means (FCM), the most well-known partitional clustering algorithm, was established in 1970's and still in use. However, there are some unsolved problems in FCM and variants of FCM are still under development. In this paper, the problems in FCM are first explained and the available solutions are investigated, which is aimed to give researchers some possible ways of future research. Most of the FCM variants try to solve the problems using domain knowledge specific to a given problem. However, in this paper, we try to give general solutions without using any domain knowledge. Although there are more things left than discovered, this paper may be a good starting point for researchers newly entered into a clustering area.

• 제1저자 : 허경용 교신저자 : 이임건
• 투고일 : 2010. 10. 07, 심사일 : 2010. 10. 27, 게재확정일 : 2010. 11. 08.
* 동의대학교 영상미디어센터(Visual Media Center, Dongeui University)
** 동의대학교 게임공학과(Dept. of Game Eng., Dongeui University)
*** 동의대학교 영상정보공학과(Dept. of Visual Information Eng., Dongeui University)
※ 이 논문은 한국콘텐츠진흥원 2010년도 문화기술 공동연구센터 사업의 지원에 의해 연구되었음.

▸ Keywords : Fuzzy c-means, Unsupervised learning, Partitional clustering, Problems in FCM

Ⅰ. 서 론

클러스터링은 주어진 데이터 집합 $X=\{x_1,x_2,\cdots,x_N\}$를 K개의 균일한(homogeneous) 부분집합으로 묶는 비교사(unsupervised) 학습 방법 중 하나로 기계 학습, 패턴 인식, 영상 분석, 생물정보학(bioinformatics), 데이터 마이닝(data mining) 등 다양한 분야에서 사용되고 있다. 클러스터링은 그 역사가 오랜 만큼 다양한 형태의 알고리듬이 제안되고 다양한 응용 분야에 맞게 변형되어 사용되고 있다[1]. 이 중 대표적인 클러스터링 기법으로는 계층적 클러스터링 (hierarchical clustering)과 분할 기반 클러스터링 (partitional clustering)이 있다. 계층적 클러스터링은 클러스터의 계층 구조를 구성하는 방식으로 하나의 클러스터에서 시작해서 연속적으로 클러스터를 나누어 가는 하향식 (top-down) 방법과 하나의 데이터 포인트로 구성되는 N개의 클러스터에서 시작해서 클러스터를 뭉쳐 가는 상향식 (bottom-up) 방법이 있다. 이에 비해 분할 기반 클러스터링은 K개의 원형(prototype)을 설정하고 가장 가까운 원형에 데이터 포인트를 할당하는 과정을 반복함으로써 K개 원형을 찾아내는 방식이다. 계층적 클러스터링이 클러스터의 분할 또는 병합 과정에서 국부적인 정보만을 활용하는 단점이 있다면 분할 기반 클러스터링은 반복 최적화 알고리듬의 사용으로 연산량의 요구가 큰 단점이 있다. 이 논문은 이들 중 가장 널리 사용되고 있는 분할 기반 클러스터링 기법을 대상으로 한다.

분할 기반 클러스터링 기법은 일반적으로 K-means[2]를 그 시초로 생각한다. K-means는 주어진 K개 원형에 데이터 포인트가 속하는지의 여부를 속하거나 (1로 표현) 속하지 않는 (0으로 표현) 이산적인 값으로 나타내므로 hard clustering이라고도 불린다. 이러한 소속 여부 표시 방법은 클러스터들이 중첩되어 나타나거나 잡음이 첨가된 경우에는 대처하기 어려우므로 소속 정도를 연속적인 소속도 값으로 나타내는 fuzzy c-means(FCM)가 제안되었다. FCM은 hard clustering에 반해 soft clustering이라고 불린다. 퍼지 소속도는 Zadeh의 퍼지 집합 이론[3]에서 연유한 것으로, Ruspini[4]가 클러스터링에 처음으로 퍼지 분할(fuzzy partition)을 소개하고 이를 이용하여 Dunn[5]이 최초의 퍼지 클러스터링 알고리듬을 제안하였다. 이후 Bezdek[6]에 의해 Dunn의 알고리듬이 일반화되어 현재 FCM은 일반적으로 Bezdek의 방법을 일컫는다.

FCM은 데이터 집합 X가 주어진 경우 K개 클러스터의 중심 V와 소속도 U가 식 (1)의 목적 함수를 최소화하도록 반복적으로 최적화한다.

$$J(V,U|X)=\sum_{i=1}^{N}\sum_{k=1}^{K}u_{ik}^{m}d_{ik}^{2} \quad \cdots\cdots (1)$$

이 때 $m(1\le m\le\infty)$은 퍼지화 정도를 나타내는 상수이며, d_{ik}는 x_i와 v_k 사이의 거리를 나타내는 척도(measure)로 FCM에서는 유클리드 거리가 사용된다. 소속도는 식 (2)의 제약 조건을 만족시켜야 하며 이는 소속도의 합이 일이 되어야(sum-to-one) 함을 나타낸다.

$$\sum_{k=1}^{K}u_{ik}=1 \qquad \forall i \quad \cdots\cdots (2)$$

식 (1)의 목적 함수는 반복 최적화 알고리듬을 이용하여 최적화하며, 라그랑즈 승수법(Lagrange multiplier method)을 이용하여 식 (3)과 (4)의 갱신 식(update equation)을 얻을 수 있다.

$$v_k=\frac{\sum_{i=1}^{N}u_{ik}^{m}x_i}{\sum_{i=1}^{N}u_{ik}^{m}} \quad \cdots\cdots (3)$$

$$u_{ik}=\left(\sum_{j=1}^{K}\left(\frac{d_{ik}}{d_{ij}}\right)^{2/(m-1)}\right)^{-1} \quad \cdots\cdots (4)$$

FCM 알고리듬은 그림 1과 같이 요약될 수 있다.

1: V를 초기화한다.
2: t를 0으로 초기화한다.
3: **do**
4: $t\leftarrow t+1$
5: 식 (4)로 U를 계산한다.
6: 식 (3)으로 V를 계산한다.
7: **while** U와 V가 수렴 조건을 만족하지 않는 경우
8: **return** U **and** V

그림 1. FCM 알고리듬
Fig. 1. FCM algorithm

FCM은 처음 소개된 이후 원형 그대로 또는 주어진 문제에 맞게 변형된 형태로 많은 문제에 성공적으로 사용되어 왔다.

하지만 많은 FCM의 변형이 존재한다는 사실은 FCM이 모든 문제에 적합한 것은 아니라는 반증이 될 수 있다. 따라서 이 논문에서는 FCM이 가지는 문제점들을 살펴보고 이를 해결할 수 있는 방법을 살펴봄으로써 향후 연구 방향을 제시한다. 기존의 많은 방법들이 문제 영역의 지식들을 활용하는 전용 방법들인데 반해 이 논문에서는 모든 문제에 적용이 가능한 범용 방법들에 중점을 둔다. 제시하는 방법들은 현재 연구가 진행 중인 방법들로 향후 많은 연구가 필요하지만 지금까지 소개된 방법들 중 성능이 우수할 뿐만이 아니라 확장 및 발전 가능성이 높은 방법들로 FCM 뿐만이 아니라 클러스터링의 문제를 해결하기에 현 시점에서 좋은 출발점이 될 것이다.

II. FCM의 문제점 및 해결 방안

FCM은 주어진 데이터 집합 X를 K개의 균일한 부분집합으로 나누는 알고리듬으로 정의할 수 있으며 FCM의 몇 가지 문제점은 이러한 FCM의 정의에서 찾을 수 있다. 먼저 FCM에서 K를 정하는 문제가 있다. FCM에서 K는 알려진 것으로 가정하지만 항상 그런 것은 아니며 대부분의 경우 시행착오를 통해 결정되므로 많은 시간을 요하는 문제가 있다. 또 다른 문제점은 FCM에서 사용하는 유클리드 거리에 있다. 유클리드 거리는 잡음에 민감하다는 사실이 널리 알려져 있다. 따라서 유클리드 거리를 사용하는 알고리듬은 잡음에 민감하며 FCM 역시 마찬가지다. 유클리드 거리에 따른 잡음 민감성은 소속도의 사용으로 일부 완화되지만 식 (2)의 제약 조건 역시 잡음 민감성의 원인이다. 이 제약 조건은 모든 소속도 값이 영이 되는 자명해(trivial solution)를 방지해 주지만 잡음에 직관적이지 않은 소속도를 할당함으로써 국부 최적해에 빠지도록 하는 원인이 된다. 유클리드 거리를 사용함으로써 발생하는 또 다른 문제점은 원형 클러스터만을 다룰 수 있다는 점이다. 마할라노비스 거리 (Mahalanovis distance)는 유클리드 거리의 단점을 보완하기 위해 많이 사용되는 일반화된 유클리드 거리이지만 이 역시 가우스 분포만을 다룰 수 있다는 단점이 있다.

그림 1에 보여진 FCM 알고리듬의 문제점은 단계 1의 초기화에 있다. 일반적으로 클러스터 중심의 초기값은 데이터 중 임의의 K개를 선택하여 사용한다. 하지만 이러한 무작위 초기화 (random initialization)는 국부 최적해만을 보장하므로 최종 클러스터링 결과를 얻기 위해서는 여러 번의 클러스터링 결과를 통해 최적의 결과를 얻는 추가적인 과정이 필요하다.

FCM의 문제점은 아니지만 클러스터링의 결과를 평가하는 방법 역시 해결해야할 과제이다. 교사 학습 (supervised learning)에서와는 다르게 비교사 학습에서는 결과를 평가하기 위해 클래스 정보를 사용할 수 없다. 따라서 클러스터링의 결과를 평가하기 위한 다양한 척도가 제안되었음에도 이는 여전히 해결해야 할 과제로 남아있다.

2.1 클러스터의 개수 K의 결정

FCM에서 클러스터의 개수는 알려진 것으로 가정하며 주어진 데이터에 적합한 클러스터의 개수는 래퍼(wrapper) 형태의 알고리듬을 써서 알아내는 것이 간단하면서도 효과적인 방법이다. 래퍼 알고리듬은 주어진 범위의 K에 대해 클러스터링을 수행하고 그 중 특정 척도를 최적화하는 K를 선택한다.

래퍼 알고리듬은 구현이 간단하고 실험적으로 우수한 성능을 보이지만 K의 범위가 알려지지 않은 경우 많은 연산을 요구하는 단점이 있다. 따라서 탐욕적인(greedy) 방법을 통해 계층적 클러스터링에서와 같이 분할 또는 병합 과정을 통해 클러스터의 개수를 증가 또는 감소시켜 나가는 방법이 많이 사용된다[7]. 탐욕적인 방법은 분할, 병합, 분할-병합의 세 부류로 나눌 수 있으며 분할과 병합을 함께 사용하는 방법이 한 가지만을 사용하는 방법에 비해 국부 최적해에 빠질 가능성이 적으며 실험적으로도 우수한 성능을 보인다.

탐욕적인 방법에서도 래퍼 알고리듬과 마찬가지로 클러스터의 분할 또는 병합을 위한 척도가 필요하다. 여기에는 클러스터링 전체를 평가하는 전역 척도와 개별 클러스터를 평가하는 국부 척도가 있다. 대표적인 전역 척도로는 BIC (Bayesian information criterion)[8]가 있으며 이는 래퍼 알고리듬에서 K를 결정하기 위해서도 사용된다. 국부 척도로는 특정 클러스터에 대한 가설 검증 (hypothesis test)이 대표적이다[9]. 가설 검증을 통한 국부 척도는 데이터의 부분집합을 이용하므로 연산량이 적고 통계학의 결과를 활용할 수 있어 휴리스틱에 의존하지 않는 등의 장점이 있다. 하지만 가우스 분포를 따르지 않는 일반적인 경우에는 사용될 수 없는 한계가 있다. 이러한 가우시안 분포의 한계는 대부분의 척도에도 적용된다.

가우시안 혼합 모델로 데이터를 한정하는 경우 K의 결정은 EM (expectation maximization)[10]의 변형을 통해서도 가능하다. EM에 분할-병합을 도입하여 가우시안 컴포넌트의 개수를 결정하는 다양한 알고리듬이 제시되어 있으며[11][12], 여기에 가설 검증을 결합함으로써 휴리스틱을 배제한 엄밀한 알고리듬을 얻을 수 있을 것이다.

2.2 유클리드 거리 사용에 따른 잡음 민감성

유클리드 거리의 잡음 민감성은 널리 알려진 사실이다. FCM 역시 유클리드 거리를 사용하므로 잡음에 민감하며 이를 해결하기 위한 방법은 크게 비-유클리드 거리를 사용하는

방법과 유클리드 거리를 사용하면서 FCM을 변형하는 두 가지가 있다. 잡음 민감성 개선을 위한 비-유클리드 거리는 일반적으로 유클리드 거리에 대한 비선형 함수로 주어지는 거리로 식 (5)와 같이 표현된다.

$$d_{ik}^{'} = f\left(\|x_i - v_k\|^2\right) \quad \cdots\cdots (5)$$

이 때 f는 임의의 단조 증가 함수이다. 이러한 유형의 거리 사용은 소속도 결정 과정에서 소속도의 비선형 변환을 통해서도 유사한 효과를 얻을 수 있으므로 여기서는 고려하지 않는다. 이외에도 다양한 비선형 거리 척도가 존재하지만 이들은 가우스안 클러스터만 다룰 수 있는 한계를 극복하기 위해 주로 사용되므로 다른 절에서 다룬다.

유클리드 거리를 사용하면서 잡음 민감성을 줄일 수 있는 방법으로는 노이즈 클러스터링(noise clustering)[13]과 regularization[14]이 있다. 노이즈 클러스터링은 가상의 노이즈 클러스터를 도입하고 이 클러스터에 소속되는 정도를 잡음의 정도로 가정함으로써 잡음의 영향을 줄인다. 하지만 모든 데이터 포인트에서 동일한 거리에 존재하는 노이즈 클러스터의 가정은 종종 원하지 않는 결과를 가져오므로 다양한 변형이 제안되었음에도 최근 관련 연구는 드문 실정이다. 이에 비해 regularization은 해공간 (solution space)에서 가능한 해의 범위를 제한함으로써 잡음의 영향을 줄인다[15].

regularization은 수학적으로 엄밀한 방법으로 그 효과가 클러스터링 이외의 영역에서도 입증되었고 실험적으로도 노이즈 클러스터링에 비해 우수한 성능을 보여주었다[14]. 비록 regularization이 클러스터링 영역에서는 최근 언급되고 있지만 문제 영역의 지식을 활용하는 이전 FCM의 변형들 중 일부도 regularization을 활용한 것으로 볼 수 있으므로 regularization 하에서 기존의 FCM 변형들을 설명하려는 시도는 잡음 민감성을 줄일 뿐만이 아니라 다양한 변형을 포괄하는 통합 모델을 제시하는 일이 될 것이다.

2.3 소속도 제약 조건에 따른 잡음 민감성

FCM에서 소속도 u_{ik}는 각 데이터 포인트 x_i가 클러스터 v_k에 소속될 정도를 [0, 1]의 범위를 가지는 연속적인 값으로 표현함으로써 부분적인 소속 정도를 표현 가능하도록 해준다. 이는 하나의 데이터 포인트가 여러 클러스터에 서로 다른 정도로 소속되는 것을 허용함으로써 잡음의 영향을 줄이는 효과가 있다. 이러한 잡음 강건성은 주성분 분석, SVM 등에 소속도를 도입한 퍼지 주성분 분석[16], 퍼지 SVM[17] 등에서 확인할 수 있다. 하지만 소속도는 식 (2)의 제약 조건을 만족시켜야 하며 이는 잡음 민감성의 또 다른 원인이다. 따라서 전통적인 소속도를 변형하여 사용하려는 다양한 시도가 있어 왔으며 이는 소속도 합이 일이 되는 제약 조건을 유지하는 방식과 제약 조건을 제거하는 방식으로 크게 나누어 볼 수 있다. 제약 조건을 만족시키면서 잡음 민감성을 줄이는 방법은 식 (5)와 유사하게 소속도를 비선형 함수를 통해 변환하는 것으로 robust statistics와 밀접한 관련이 있으며 특히 M-estimator를 이용한 방법이 주로 사용된다[18].

소속도 합이 일이 되는 제약 조건을 제거한 대표적인 방법으로는 전형도 (typicality)를 이용한 possibilistic c-means (PCM)가 있다[19]. 이 역시 M-estimator의 일종으로 볼 수 있으며 제약 조건이 없음으로 발생하는 자명해는 regularization을 통해 해결하고 있다. 하지만 PCM은 초기 조건에 지나치게 민감하기 때문에 초기값 설정에 주의하여야 한다. 이러한 단점을 보완하기 위해 FCM과 PCM을 결합한 possibilistic fuzzy c-means (PFCM)[20]는 소속도와 전형도를 동시에 사용함으로써 상호 단점을 보완하여 현재까지 알려진 퍼지 클러스터링 알고리듬 중 가장 나은 성능을 보이는 알고리듬 중 하나이다. PFCM 역시 마할라노비스 거리의 도입[21], regularization의 도입[14] 등을 통해 다양한 변형이 시도되고 있으며 robust statistics의 도입은 소속도 문제 개선의 출발점이 될 것이다.

2.4 가우스 분포 제한

유클리드 또는 마할라노비스 거리를 유사도로 사용하는 경우의 문제점 중 하나는 가우스 혼합 모델에서만 사용할 수 있다는 점이다. 비록 모든 데이터가 가우스 혼합 모델로 근사화될 수 있는 것으로 알려져 있지만 실세계의 데이터는 제한된 데이터 크기로 인해 이 조건을 만족시키는 경우가 드물다. 따라서 가우스 혼합 모델을 만족시키지 않는 데이터에 FCM을 적용하려는 다양한 변형들이 제안되었다. 이들 방법은 특징 공간 (feature space)에서 비-유클리드 거리를 이용하는 방법과 커널을 이용하여 커널 특징 공간 (kernel feature space)으로 사상한 후 유클리드 거리를 이용하는 방법으로 크게 나누어볼 수 있다. 대표적인 비-유클리드 거리로는 geodesic distance[22]와 random walk distance[23]가 있다. geodesic distance는 구체에서 표면 거리를 측정하는 방법에서 유래된 거리로 오목한 클러스터들도 다룰 수 있는 장점이 있지만 잡음에 민감한 것으로 알려져 있다. 이에 비해 random walk distance는 그래프 이론에서 연유한 최단 거리의 변형으로 잡음에 강하다.

커널을 이용하는 방법은 SVM[24]의 성공 이후 널리 적용되고 있는 커널 이론을 클러스터링을 적용한 것으로 Girolami[25]

에 의해 처음 클러스터링에 도입된 이후 다양한 방법들이 제시되고 사용되었다[26]. 커널 이론은 커널 특징 공간으로의 사상 후 선형 변환을 수행하는 것은 특징 공간에서의 비선형 변환에 해당한다는 것이 기본 개념이다. 커널 클러스터링은 사용하는 커널의 종류에 따라 다양한 효과를 가져올 수 있으며 선택할 수 있는 커널의 종류 또한 다양하므로 확장성과 응용력이 뛰어나 비-유클리드 거리를 이용하는 방법에 비해 많이 사용된다. 하지만 아직은 커널의 종류에 따른 피처 공간에서의 효과가 명확하게 밝혀지지 않았고 대용량의 데이터를 다루기에 적합하지 않은 문제점 등이 있어 지금도 많은 연구가 진행되고 있는 분야 중 하나이다.

또 다른 비선형 클러스터링 기법으로는 spectral clustering이 있다[27][28]. spectral clustering은 그래프 이론에서 연유한 방법으로 반복 최적화 기법을 사용하는 FCM 류의 방법과 다르게 국부 최적해가 존재하지 않으며 고유치 문제를 통해 클러스터링 문제를 해결할 수 있는 등의 장점으로 인해 널리 알려진 방법이다. spectral clustering에서 풀어야할 가장 큰 문제는 유사도 행렬을 구성하는 방법으로 유사도 행렬이 가져야하는 바람직한 특징은 알려져 있다[29]. 하지만 알려진 특징을 만족시키는 유사도 행렬을 구하는 과정은 여전히 시행착오에 의존하고 있으며 특정 조건 하에서 커널 클러스터링과 동등한 것으로 알려지면서[30] 현재로는 커널 클러스터링에 비해 많이 연구되고 있지는 않다.

2.5 클러스터 중심의 초기값 설정

FCM은 반복 최적화 기법을 통해 최적해를 찾는다. 하지만 기울기 하강(gradient descent) 방법을 사용하는 최적화 기법의 문제점은 국부 최적해에 빠진다는 점이다. 특히 FCM에서 얻어지는 국부 최적해는 클러스터 중심의 초기값에 영향을 받는다. 전역 최적해를 얻을 수 있는 초기값을 구하는 문제는 NP-complete로 알려져 있으므로[31] 유사 최적해를 구하기 위한 많은 시도가 있어왔다. 이들은 크게 무작위 샘플링(random sampling)을 사용하는 방법, 거리 최적화 방법, 밀도 추정 방법의 세 가지로 나눌 수 있다[32]. 무작위 샘플링은 간단하지만 많은 연산을 요하는 단점이 있다. 거리 최적화는 FCM에서 간과하는 클러스터들 사이의 거리를 이용하는 장점이 있지만 FCM의 수행 과정에서 이 정보가 사용되지 않으므로 단점이 될 수 있다. 밀도 추정 방법은 밀집된 샘플 공간에 클러스터 중심의 초기값을 두는 방법으로 이 중 전역 클러스터링 (global clustering)은 특정 목적 함수를 최적화하는 데이터 포인트를 반복적으로 초기값에 추가해 나가는 결정적인 (deterministic) 알고리듬이다[33][34]. 전역 FCM (GFCM)에서는 식 (1)의 목적 함수를 최적화하는 데이터 포인트를 초기값으로 선택함으로써 FCM과의 연관성이 높은 점은 또 다른 장점이다. 비록 GFCM의 성능이 다른 방법들에 비해 탁월하다고는 할 수 없지만 초기값 선택을 위한 목적 함수의 변형을 통해 여러 가지 효과를 얻을 수 있는 장점이 있으며 다른 초기화 방법을 아우르는 일반화된 방법으로의 발전 가능성이 있는 방법이다.

2.6 클러스터링 평가 방법 및 기타

클러스터링은 대표적인 비교사 학습법 중 하나로 클래스 표시자와 같은 학습의 목표값이 존재하지 않으므로 클러스터링 결과를 비교하는 것은 쉬운 일이 아니다. 간단하면서도 가장 널리 쓰이는 클러스터링 평가 기준은 “클러스터의 밀집도는 높고 클러스터들은 서로 멀리 떨어져 존재한다”는 것이다. 이러한 기준을 사용하는 대표적인 척도 중 하나가 Xie-Beni의 척도로[35] 제안된 많은 척도들 중 가장 나은 성능을 보인다. 하지만 이 척도 역시 직관에 반하는 결과를 보여주는 경우가 있다.

클러스터링을 평가하는 다른 방법으로는 정보 이론(information theory)를 활용하는 방법이다[36][37]. 정보 이론의 활용을 통한 클러스터링의 평가는 휴리스틱에 의존하지 않고 이론적인 배경을 얻을 수 있는 장점으로 최근 관련 연구가 증가하고 있다. 하지만 클러스터링의 결과는 이후 분류(classification) 등에서 사용되는 경우가 많으므로, 이러한 경우에는 분류 오류를 통해 판단하는 것이 쉽고 바람직하다. 클러스터링 결과를 평가해야만 하는 경우에는 여러 척도들을 동시에 사용하고 이들 중 최적 결과의 선택 또는 통합(aggregation)을 통해 최적의 클러스터링을 결정하는 것이 바람직하다.

이외에도 고차원의 데이터를 다루는 방법, 대량의 데이터를 다루는 방법 등이 최근 주목을 받고 있다. 대표적인 고차원의 데이터로는 이미지 데이터[38], 유전자 데이터[39] 등이 있다. 하지만 고차원의 데이터는 일반적으로 차원 축소[40]와 같은 전처리 과정을 도입하고 이후 저차원의 데이터에 대해 클러스터링을 수행하는 것이 효율적이다. 또한 대량의 데이터를 다루는 방법은 구현의 측면이 강조되므로 이 논문에서는 다루지 않는다.

III. 결론

클러스터링은 주어진 데이터를 균일한 부분집합으로 나누

는 비교사 학습법 중 하나로 특히 FCM은 1970년대 정립된 이후 지금까지도 널리 사용되는 대표적인 클러스터링 알고리듬 중 하나이다. 하지만 FCM은 여러 가지 해결되지 못한 문제들이 있으며 이 논문에서는 이들 문제점들을 살펴보고 이를 해결할 수 있는 연구 방향을 살펴 보았다. 비록 제시된 방법들이 기존 방법들에 비해 항상 나은 성능을 보이지는 않으며 향후 더 많은 연구가 필요한 방법들이지만 이들 사이에는 한 가지 공통점을 발견할 수 있다. 즉, 수학 또는 통계학에서 정립된 이론들을 도입하여 FCM의 문제점을 해결하고자 하는 시도들이 다른 방법들에 비해 활용도 및 확장 가능성이 높다는 점이다. FCM의 문제점을 해결하기 위해 도입된 수학 및 통계학 이론들로는 가설 검증, regularization, robust statistics, 커널 이론, 그래프 이론, 정보 이론 등이 있다. 이러한 이론을 이용한 것으로 명시적으로 언급하지는 않았지만 많은 FCM의 변형들이 기존 이론의 틀에서 설명될 수 있는 것을 볼 때, 향후 FCM의 연구 방향은 기존 이론의 적용 범위를 명확히 하는 일반화된 알고리듬으로의 발전과 클러스터링이라는 주제에 맞게 이론을 재정립하는 방향으로 진행될 것으로 생각된다. 또한 이 논문에서는 다루지 않았지만 FCM의 개선에 더불어 주어진 문제의 사전 지식을 활용함으로써 보다 나은 성능을 얻을 수 있을 것이다.

참고문헌

[1] R. Xu and D. Wunsch, Clustering, Wiley-IEEE Press, 2008.

[2] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, pp. 281-297, 1967.

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

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

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

[6] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Springer, 1981.

[7] H. Frigui and R. Krishnapuram, "Clustering by competitive agglomeration," Pattern Recognition, vol. 30, no. 7, pp. 1109-1119, 1997.

[8] Gyeongyong Heo, Young Woon Woo, "Extensions of X-means with Efficient Learning the Number of Clusters ," Journal of the KIMICS, Vol. 12, No. 4, pp. 772-780, 2008

[9] G. Heo and P. Gader, "Learning the Number of Gaussian Components Using Hypothesis Test," Proceedings of the 2009 International Joint Conference on Neural Networks, pp. 1206-1212, 2009.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Series B, vol. 39, no. 1, pp. 1-38, 1977.

[11] Z. Zhang, C. Chen, J. Sun, and K. L. Chan, "EM algorithms for Gaussian mixtures with split-and-merge operation," Pattern Recognition, vol. 36, no. 9, pp. 1973-1983, 2003.

[12] Y. Li and L. Li, "A Novel Split and Merge EM Algorithm for Gaussian Mixture Model," Proceedings of the 5th International Conference on Natural Computation, pp. 479-483, 2009.

[13] R. N. Dave, "Characterization and detection of noise in clustering," Pattern Recognition Letters, vol. 12, no. 11, pp. 657-664, 1991.

[14] Y. Namkoong, G. Heo, and Y. W. Woo, "An Extension of Possibilistic Fuzzy C-Means with Regularization," Proceedings of the 2010 IEEE International Conference on Fuzzy Systems, pp. 696-701, 2010.

[15] A. Tikhonov, "On solving incorrectly posed problems and method of regularization," Dokl. Acad. Nauk USSR, vol. 151, pp. 501-504, 1963.

[16] G. Heo, P. Gader, and H. Frigui, "RKF-PCA: Robust Kernel Fuzzy PCA," Neural Networks, vol. 22, no. 5-6, pp. 642-650, 2009.

[17] C. F. Lin and S. D. Wang, "Fuzzy support vector machines," IEEE Transactions on Neural Networks, vol. 13, no. 2, pp. 464-471, 2002.

[18] P. J. Huber and E. M. Ronchetti, Robust Statistics, 2nd edition, Wiley, 2009.

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

[20] 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.

[21] Gyeongyong Heo, Sewoon Choe, Young Woon Woo, " Improvement of the PFCM(Possibilistic Fuzzy C-Means) Clustering Method," Journal of the KIMICS, Vol. 13, No. 1, pp. 177-185, 2009.

[22] B. Feil and J. Abonyi, "Geodesic Distance Based Fuzzy Clustering," Advances in Soft Computing, vol. 39/2007, pp. 50-59, 2007.

[23] F. Fouss, A. Pirotte, J. M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3, pp. 355 - 369, 2007.

[24] B. Scholkopf and A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, 2001.

[25] M. Girolami, "Mercer kernel-based clustering in feature space," IEEE Transactions on Neural Networks, vol. 13, no. 3, pp. 780 - 784, 2002.

[26] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," Pattern Recognition, vol. 41, no. 1, pp. 176 - 190, 2008.

[27] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, no. 8, pp. 888 - 905, 2000.

[28] U. von Luxburg, "A tutorial on spectral clustering," Statistics and Computing, vol. 17, no. 4, pp. 395 - 416, 2007.

[29] Gyeongyong Heo, Kwang-Baek Kim, Young Woon Woo, "Magnifying Block Diagonal Structure for Spectral Clustering ," Journal of Korea Multimedia Society, Vol. 11, No. 9, pp. 1302-1309, 2008

[30] I. S. Dhillon, Y. Guan, and B. Kulis, "A unified view of kernel k-means, spectral clustering and graph cuts," Department of Computer Science, University of Texas, Tech. Rep. TR-04-25, 2005.

[31] M. Garey, D. Johnson, and H. Witsenhausen, "The complexity of the generalized Lloyd-Max problem," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 255-256, 1982.

[32] J. He, M. Lan, C. L. Tan, S. Y. Sung, and H. B. Low, "Initialization of cluster refinement algorithms: A review and comparative study," Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, pp. 297 - 302, 2004.

[33] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," Pattern Recognition, vol. 36, pp. 451 - 461, 2003.

[34] G. Heo and P. Gader, "An Extension of Global Fuzzy C-means Using Kernel Methods," Proceedings of the 2010 IEEE International Conference on Fuzzy Systems, pp. 690-695, 2010.

[35] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 8, pp. 841-847, 1991.

[36] M. Meila, "Comparing clusterings - an information based distance," Journal of Multivariate Analysis, vol. 98, no. 5, pp. 873 - 895, 2007.

[37] D. Pascual, F. Pla, and J. S. Sanchez, "Cluster validation using information stability measures," Pattern Recognition Letters, vol. 31, pp. 454-461, 2010.

[38] Q. Deng, Y. Luo, and J. Ge, "Dual threshold based unsupervised face image clustering," Proceedings of the 2nd International Conference on Industrial Mechatronics and Automation, pp. 436-439, 2010.

[39] D. Jiang, C. Tang, A. Zhang, "Cluster analysis for gene expression data: a survey," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1370-1386, 2004.

[40] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality Reduction: A Comparative Review," Tilburg University, Technical Report, TiCC-TR 2009-005, 2009.

저 자 소 개

허 경 용

1996년 8월 : 연세대학교 본대학원 전자공학과 (공학석사)
2009년 12월 : Department of Computer and Information Science and Engineering, University of Florida (공학박사)
관심분야 : Machine Learning, Pattern Recognition, Image Processing
Email : hgycap@hotmail.com

서 진 석

1998년 2월 : 건국대학교 공학사
2000년 2월 : 포스텍 공학석사
2005년 2월 : 포스텍 공학박사
2005년-현재 : 동의대학교 게임공학과 조교수
관심분야 : 컴퓨터 게임, 저작도구, 가상현실, 증강현실
Email : jsseo@deu.ac.kr

이 임 건

1991년 : 연세대학교 공학사
1993년 : 연세대학교 공학석사
1998년 : 연세대학교 공학박사
2002년 - 현재 : 동의대학교 영상 정보공학과 교수
관심분야 : 영상복원, 영상신호처리 컴퓨터비전
Email : iglee@deu.ac.kr