Regularization을 이용한 Possibilistic Fuzzy C-means의 확장

An Extension of Possibilistic Fuzzy C-means using Regularization

저자
(Authors)

허경용, 남궁영환, 김성훈
Gyeong-Yong Heo, Young-Hwan NamKoong, Seong-Hoon Kim

출처
(Source)

한국컴퓨터정보학회논문지 15(1), 2010.1, 43-50(8 pages)
Journal of the Korea Society of Computer and Information 15(1), 2010.1, 43-50(8 pages)

발행처
(Publisher)

한국컴퓨터정보학회
The Korean Society Of Computer And Information

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06528077

APA Style

허경용, 남궁영환, 김성훈 (2010). Regularization을 이용한 Possibilistic Fuzzy C-means의 확장. 한국컴퓨터정보학회논문지 , 15(1), 43-50

韓國컴퓨터情報學會 論文誌
第15卷 第1號, 2010. 1.

2010-15-1-1-5

Regularization을 이용한 Possibilistic Fuzzy C-means의 확장

허 경 용*, 남 궁 영 환*, 김 성 훈**

An Extension of Possibilistic Fuzzy C-means using Regularization

Gyeongyong Heo*, Younghwan Namkoong*, Seong Hoon Kim**

요 약

Fuzzy c-means(FCM)와 possibilistic c-means(PCM)는 퍼지 클러스터링 영역에서 대표적인 두 가지 방법으로 많은 패턴 인식 문제들에 성공적으로 활용되어져 왔다. 하지만 이들 방법 역시 잡음 민감성과 중첩 클러스터 문제를 가지고 있다. 이들 문제점을 극복하기 위해, 최근 두 방법을 결합하려는 시도가 있어왔고, possibilistic fuzzy c-means(PFCM)는 FCM과 PCM을 목적 함수 단계에서 통합함으로써 두 방법이 가지는 문제점을 완화시키는 성공적인 결과를 보여주었다. 이 논문에서는 PFCM에 regularization을 도입함으로써 PFCM의 잡음 민감성을 한층 더 줄여줄 수 있는 향상된 PFCM을 소개한다. Regularization은 해공간을 평탄화 함으로써 잡음의 영향을 줄이는 대표적인 방법 중 하나이다. 제안한 방법은 PFCM의 장점과 더불어 regularization에 의해 잡음의 영향을 더욱 줄일 수 있으며, 이는 실험을 통해 확인할 수 있다.

Abstract

Fuzzy c-means (FCM) and possibilistic c-means (PCM) are the two most well-known clustering algorithms in fuzzy clustering area, and have been applied in many applications in their original or modified forms. However, FCM's noise sensitivity problem and PCM's overlapping cluster problem are also well known. Recently there have been several attempts to combine both of them to mitigate the problems and possibilistic fuzzy c-means (PFCM) showed promising results. In this paper, we proposed a modified PFCM using regularization to reduce noise sensitivity in PFCM further. Regularization is a well-known technique to make a solution space smooth and an algorithm noise insensitive. The proposed algorithm, PFCM with regularization (PFCM-R), can take advantage of regularization and further reduce the effect of noise. Experimental results are given and show that the proposed method is better than the existing methods in noisy conditions.

▸ Keyword : 퍼지 클러스터링 (Fuzzy Clustering), Possibilistic Fuzzy C-means (Possibilistic Fuzzy C-means), 잡음 민감성 (Noise Sensitivity), Regularization (Regularization)

• 제1저자 : 허경용 교신저자 : 김성훈
• 투고일 : 2010. 01. 05, 심사일 : 2010. 01. 11, 게재확정일 : 2010. 01. 26.
* Computer and Information Science and Engineering, University of Florida ** 경북대학교 컴퓨터정보학부 교수

Ⅰ. 서 론

클러스터링은 주어진 데이터를 유사성에 기준하여 몇 개의 그룹으로 나누는 방법으로 패턴 인식의 주요 기법 중 하나이다. 소속도 함수(membership function)에 의해 부분 소속도를 나타내는 퍼지 집합이 Zadeh에 의해 소개된 이후[1], 퍼지 집합은 클러스터링 분야에 도입되었고 퍼지 클러스터링은 대표적인 클러스터링 기법 중 하나로 자리 잡았다. Bezdek[2]에 의해 일반화된 fuzzy c-means(FCM)는 퍼지 클러스터링 방법 중 대표적인 방법이다. FCM은 간단하면서도 효과적인 클러스터링 방법이지만, 구해진 소속도가 직관적인 값과 일치하지 않는 경우가 있으며 잡음이 많은 환경에서는 정확한 소속도를 구할 수 없는 문제점이 있다. 이러한 문제점의 원인 중 하나는 소속도 값의 합이 1이 되어야한다는 제약 조건(sum-to-one constraint)이다. 유사도를 판단하기 위해 사용하는 거리 척도(distance measure) 역시 그 원인 중 하나이지만, regularization은 모든 거리 척도와 함께 사용할 수 있으므로, 이 논문에서는 소속도의 제약 조건에 따른 잡음 민감성 개선을 목표로 하며, 거리 척도는 다루지 않는다. FCM의 잡음 민감성을 해결하기 위한 방법 중 한 가지는 Krishnapuram 등이 제안한 possibilistic c-means(PCM)이다[3]. PCM은 소속도가 아닌 전형도(typicality)를 사용하며 이는 FCM에서와 같은 제약 조건을 가지지 않으므로 잡음 민감성을 줄일 수 있다. 소속도는 어떤 데이터가 특정 클러스터에 속할 상대적인 값인데 비해 전형도는 절대적인 값이라는 점에서 서로 다르다. 비록 이러한 전형도의 특징이 FCM의 잡음 민감성을 줄일 수 있도록 해주지만, PCM의 중첩 클러스터 문제를 야기하는 원인이기도 하다. 이처럼 FCM과 PCM은 각각의 장단점을 가지고 있으므로, 이 두 방법을 결합하여 상호 보완하려는 시도가 있어왔고[4][5][6], 그 중, Pal 등이 제안한 possibilistic fuzzy c-means(PFCM)는 두 방법을 목적 함수(objective function) 단계에서 결합함으로써 FCM의 잡음 민감성을 줄이고 PCM의 중첩 클러스터 문제를 해결하고자 하였다. PFCM은 다른 방법에 비해 나은 결과를 보여주었지만 목적 함수가 복잡해짐에 따라 안정된 해를 구하기 위해서는 보다 많은 데이터가 필요하며 해공간(solution space)에 많은 국부 최적해를 생성하는 문제점이 있다. 따라서 이 논문에서는 PFCM에 regularization을 도입하여 PFCM의 잡음 민감성을 한층 더 줄이는 방법을 제안한다. Regularization은 해공간을 평탄화함으로써 잡음에 민감하지 않은 유사해를 구하도록 해주는 방법으로 데이터가 적은 경우에도 사용할 수 있다. 제안한 방법의 유효성은 실험 결과를 통해서 확인할 수 있다.

이 논문의 구성은 다음과 같다. 먼저 2장에서는 기존 퍼지 클러스터링 방법들과 그 문제점들을 보이며 3장에서 regularization을 소개한다. 이 논문에서 제안한 방법, PFCM과 regularization을 이용하여 잡음 민감성을 줄이는 방법은 4장에서 설명하며, 5장에서는 실험을 통해 기존의 방법에 비해 제안한 방법의 잡음 민감성이 줄어드는 것을 보인다. 결론 및 향후 연구 방향에 대해서는 6장에서 언급한다.

Ⅱ. 퍼지 클러스터링

퍼지 클러스터링은 제약 조건이 있는 최적화 문제(constrained optimization problem)로, 식 (1)의 목적 함수를 최적화하는 것으로 볼 수 있다[2].

$$J_{FCM} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} D_{ik}^{2} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} \|v_i - x_k\|^2 \quad \cdots (1)$$

이 때 $1 < m < \infty$은 퍼지화 정도를 나타내는 상수로 일반적으로 2로 설정되며, n은 데이터 포인트의 개수를, c는 클러스터의 개수를, v_i는 i번째 클러스터의 중심을, u_{ik}는 k번째 데이터 포인트가 i번째 클러스터에 소속되는 정도를, 그리고 D_{ik}는 k번째 데이터 포인트와 i번째 클러스터 중심 사이의 거리를 나타낸다. 이 논문에서는 가우시안 분포를 나타낼 수 있는 Mahalanobis 거리 척도를 사용하였다. FCM의 잡음 민감성은 소속도에 주어지는 제약 조건, 소속도의 합이 1이 되어야한다는 조건에 기인한다.

$$\sum_{i=1}^{c} u_{ik} = 1 \quad \cdots (2)$$

FCM의 잡음 민감성을 줄이기 위해 PCM은 소속도가 아닌 전형도를 사용한다. 전형도는 식 (2)의 제약 조건을 제거함으로써 FCM의 잡음 민감성을 효과적으로 줄일 수 있음이 입증되었다[3]. PCM의 목적 함수는 식 (3)과 같다.

$$J_{PCM} = \sum_{k=1}^{n}\sum_{i=1}^{c} t_{ik}^{\eta} D_{ik}^{2} + \sum_{i=1}^{c}\gamma_i \sum_{k=1}^{n} (1 - t_{ik})^{\eta} \quad \cdots (3)$$

이 때 t_{ik}는 k번째 데이터 포인트가 i번째 클러스터에 소속되는 전형도를 나타내며, η는 전형도의 퍼지화 정도를 나타

내는 상수로 소속도의 퍼지화 정도를 나타내는 m과 같은 2로 설정하였다. γ_i는 클러스터의 부피를 나타내는 값으로 일반적으로 식 (4)와 같이 정의된다.

$$\gamma_i = \frac{\sum_{k=1}^{n} t_{ik}^{\eta} D_{ik}^2}{\sum_{k=1}^{n} t_{ik}^{\eta}} \quad \cdots\cdots (4)$$

PCM은 sum-to-one 조건을 제거함으로써 잡음 민감성을 줄일 수 있지만, 각 클러스터가 다른 클러스터들에 영향을 받지 않는 독립성으로 인해 중첩 클러스터 문제를 야기한다. 식 (3)은 식 (5)와 같이 서로 독립적인 c개 값들의 합으로 나타낼 수 있다.

$$J_{PCM} = \sum_{i=1}^{c}\left(\sum_{k=1}^{n} t_{ik}^{\eta} D_{ik}^2 + \gamma_i \sum_{k=1}^{n} (1-t_{ik})^{\eta}\right) \quad \cdots\cdots (5)$$

이러한 독립성은 PCM이 중첩된 클러스터 또는 동일한 클러스터를 찾는 경우가 발생하도록 한다. 이러한 FCM과 PCM의 단점을 극복하기 위해 다양한 시도가 있어왔고, 두 방법의 목적 함수를 결합하는 PFCM이 그 중 하나이다. PFCM은 FCM과 PCM의 장점을 통해 서로의 단점을 보완하는 성공적인 결과를 보여주었다. PFCM의 목적 함수는 식 (6)과 같다.

$$J_{PFCM} = \sum_{k=1}^{n}\sum_{i=1}^{c}(au_{ik}^{m} + bt_{ik}^{\eta}) D_{ik}^2 + \sum_{i=1}^{c}\gamma_i \sum_{k=1}^{n}(1-t_{ik})^{\eta} \quad \cdots\cdots (6)$$

이 때 a와 b는 소속도와 전형도의 가중치를 정해주는 상수이다. 식 (6)은 alternating optimization 기법으로 최적화할 수 있으며 소속도, 전형도, 클러스터 중심의 update equation은 식 (7), (8), (9)와 같다.

$$u_{ik} = \left(\sum_{j=1}^{c}\left(\frac{D_{ik}}{D_{jk}}\right)^{2/(m-1)}\right)^{-1} \quad \cdots\cdots (7)$$

$$t_{ik} = \frac{1}{1+\left(\frac{b}{\eta_i} D_{ik}^2\right)^{1/(\eta-1)}} \quad \cdots\cdots (8)$$

$$v_i = \frac{\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta}) x_k}{\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta})} \quad \cdots\cdots (9)$$

비록 PFCM이 PCM의 전형도를 통해 잡음 민감도를 줄이고 FCM의 sum-to-one 제약 조건을 통해 중첩 클러스터 문제를 완화하였지만, 여전히 잡음의 영향을 받는다. PFCM의 잡음 민감도 문제는 여러 방법으로 완화될 수 있으며, 이 논문에서는 regularization을 이용하여 잡음 민감도를 줄이는 방법을 제안한다.

III. Regularization

많은 패턴 인식 문제들은 식 (10)과 같은 선형 방정식을 푸는 문제로 변환될 수 있다.

$$Az = y \quad \cdots\cdots (10)$$

이 때 A는 연산자를 나타내고, z는 해를 구하고자 하는 변수를, 그리고 y는 입력 변수를 나타낸다. 식 (10)과 같이 표현되는 문제는 행렬 A의 역행렬을 통해 해를 구할 수도 있지만, 항상 역행렬을 구할 수 있는 것은 아니며, 구한다 하더라도 주어진 문제의 특성으로 인해 무의미한 해가 나올 수 있다. 이러한 문제점을 해결하기 위해 Tikhonov는 해공간을 제약 조건을 사용하여 제한하는 regularization을 제안하였다[7]. Regularization을 이용한 근사해는 식 (11)의 최적화 문제로 나타내어질 수 있다.

$$\min_z \|Az-y\| + \beta\Phi(z) \quad \cdots\cdots (11)$$

이 때 β는 regularization 상수를 나타내고 Φ는 제약 조건을 나타낸다. 제약 조건은 가능한 해의 범위를 줄이는 의미가 있지만, 제약 조건을 선택하는 일반적인 방법은 알려져 있지 않으며, 일반적으로 z의 도함수가 많이 사용된다. Regularization은 클러스터링 문제에도 적용되어 왔으며 몇 가지 목적 함수들이 제안되었다[8][9][10][11].

$$J_{Entropy} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}\|x_k - v_i\|^2 + \beta\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}\log u_{ik} \quad \cdots\cdots (12)$$

$$J_{Quadratic} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}||x_k - v_i||^2 + \beta \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2 \quad \cdots\cdots (13)$$

$$J_{Polynomial} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m||x_k - v_i||^2 + \beta \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m \quad \cdots\cdots (14)$$

식 (12), (13), (14)에서 regularization 항은 서로 약간의 차이가 있지만, 모든 항들은 u_{ik}가 동일한 값, $1/c$를 가질 때 최소가 된다. 즉, 모든 regularization 항은 u_{ik}가 0이나 1의 극단치(extreme value)를 가지지 않도록 함으로써 해공간을 평탄화하고 잡음 민감성을 줄여준다. 위의 목적 함수들은 모두 FCM에 regularization을 적용한 경우이며 PFCM에 regularization을 적용한 예는 아직 알려진 바 없다.

IV. Regularization을 이용한 PFCM의 확장

이 장에서는 regularization을 이용하여 PFCM의 잡음 민감성을 줄여주는 PFCM-R을 소개한다. 엔트로피 항이나 2차 항을 사용하는 경우, PFCM-R의 update equation을 폐쇄형(closed-form)으로 얻어낼 수 없으므로 이 논문에서는 식 (14)의 다항식 항을 사용하였다. 또한 이 논문에서는 소속도에만 regularization을 적용하고 전형도에는 적용하지 않았다. 이는 PCM의 목적 함수에 regularization 항과 유사한 기능을 하는 항이 이미 존재하기 때문이다. 식 (3)의 두 번째 항이 바로 그 항으로, 이 항은 전형도가 영의 값을 가지는 자명해(trivial solution)를 방지하기 위해 사용되었다. 따라서 전형도에 대한 regularization은 고려하지 않았다. 마지막으로, 식 (6)의 가중치 상수 a와 b는 관례에 따라 1로 설정하였다[4]. 위의 모든 내용을 고려한 PFCM-R의 목적 함수는 식 (15)와 같다.

$$J = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik}^m + t_{ik}^\eta)D_{ik}^2 + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1-t_{ik})^\eta + \beta\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^m \quad \cdots\cdots (15)$$

PFCM-R의 제약 조건은 식 (16)과 같다.

$$0 \le u_{ik} \le 1,\ \sum_{i=1}^{c} u_{ik} = 1,\ \sum_{k=1}^{n} u_{ik} > 0 \quad \cdots\cdots (16a)$$

$$0 \le t_{ik} \le 1,\ \sum_{k=1}^{n} t_{ik} > 0 \quad \cdots\cdots (16b)$$

$$m, \eta > 1 \quad \cdots\cdots (16c)$$

$$\beta > 0 \quad \cdots\cdots (16d)$$

식 (15)의 목적 함수와 식 (16)의 제약 조건을 이용하여 라그랑지 방정식(Lagrange equation)을 식 (17)과 같이 나타낼 수 있다.

$$L = \sum_{k=1}^{n}\sum_{i=1}^{c}(au_{ik}^m + bt_{ik}^\eta)D_{ik}^2 + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1-t_{ik})^\eta + \beta\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^m - \sum_{k=1}^{n}\lambda_k\left(\sum_{i=1}^{c}u_{ik} - 1\right) \quad \cdots\cdots (17)$$

이 때 $\lambda = [\lambda_1, \cdots, \lambda_k]$는 라그랑지 상수 벡터를 나타낸다. 식 (17)을 u_{ik}에 대해 편미분하면 식 (18)을 얻을 수 있다.

$$\frac{\partial L}{\partial u_{ik}} = \mu_{ik}^{m-1}D_{ik}^2 + \beta\mu_{ik}^{m-1} - \lambda_k = 0 \quad \cdots\cdots (18)$$

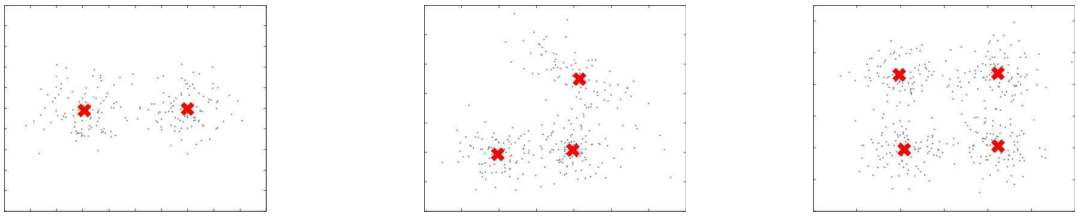
식 (18)을 u_{ik}에 대해 정리하면 식 (19)를 얻을 수 있고,

$$u_{ik} = \left(\frac{\lambda_k/m}{D_{ik}^2 + \beta}\right)^{1/(m-1)} = \frac{\lambda'_k}{D_{ik}^2 + \beta^{1/(m-1)}} \quad \cdots\cdots (19)$$

식 (19)에서 상수 λ'_k는 sum-to-one 조건을 이용하여 식 (20)과 같이 구할 수 있다.

$$\sum_{i=1}^{c} u_{ik} = \lambda'_k\sum_{i=1}^{c}(D_{ik}^2 + \beta)^{-1/(m-1)} = 1 \quad \cdots\cdots (20a)$$

$$\lambda'_k = \frac{1}{\sum_{i=1}^{c}(D_{ik}^2 + \beta)^{-1/(m-1)}} \quad \cdots\cdots (20b)$$



(a) D_2 (b) D_3 (c) D_4

그림 1. 데이터 집합
Fig. 1. Data sets

식 (19)와 (20)을 이용하여 u_{ik}의 update equation은 식 (21)과 같이 나타낼 수 있다.

$$u_{ik} = \lambda'_k (D_{ik}^2 + \beta)^{-1/(m-1)} \quad \cdots\cdots (21)$$
$$= \frac{(D_{ik}^2 + \beta)^{-1/(m-1)}}{\sum_{j=1}^{c} (D_{jk}^2 + \beta)^{-1/(m-1)}}$$
$$= \left(\sum_{i=1}^{c} \frac{(D_{ik}^2 + \beta)^{1/(m-1)}}{(D_{jk}^2 + \beta)^{1/(m-1)}} \right)^{-1}$$

식 (21)은 식 (7)과 유사한 형태를 가지지만 regularization 상수 β가 추가된 점이 다르다. 하지만 regularization을 사용하지 않는 경우, 즉, $\beta = 0$인 경우 식 (21)은 식 (7)과 동일하다. 또한 식 (21)은 식 (7)과는 다른 거리 척도를 사용한 것으로도 볼 수 있다. $D_{ik}^{'2}$을 식 (22)와 같이 정의하면,

$$D_{ik}^{'2} = D_{ik}^2 + \beta \quad \cdots\cdots (22)$$

식 (21)은 식 (23)과 같이 나타낼 수 있다.

$$u_{ik} = \left(\sum_{i=1}^{c} \frac{(D_{ik}^{'2})^{1/(m-1)}}{(D_{jk}^{'2})^{1/(m-1)}} \right)^{-1} \quad \cdots\cdots (23)$$

식 (23)은 식 (7)과 동일한 형태이다. Regularization 상수 β는 데이터 포인트와 클러스터 중심 사이의 최소 거리를 제한함으로써 소속도가 0이나 1의 극단치를 가지는 것을 방지한다. β 값이 큰 경우, u_{ik}는 $1/c$의 동일한 값을 가지게 되어 regularization 항이 소속도 결정의 주된 역할을 하게 되고, β 값이 작은 경우에는 PFCM-R이 PFCM과 같아지도록 하여 잡음 민감도가 증가한다. 따라서 β 값은 주어진 데이터에 맞게 결정되어야 한다. β 값을 결정하기 위해 수많은 방법이 제시되었지만, 아직 모든 문제에서 사용할 수 있는 방법은 존재하지 않는다[12][13]. 이 논문에서는 간단하면서도 가장 널리 사용되는 격자 탐색법(grid search method)을 이용하여 β 값을 결정하였다[14]. PFCM-R에서 전형도와 클러스터 중심에 대한 update equation은 식 (8) 및 (9)와 같다.

V. 실험 결과 및 고찰

제안한 방법이 기존 방법에 비해 나은 결과를 보인다는 것을 확인하기 위해, PFCM과 PFCM-R을 Matlab으로 구현하고 그림 1에 보인 3가지 데이터 집합에 대해 두 가지 종류의 잡음과 다양한 잡음 비율의 환경에서 실험하였다. 잡음은 일반적으로 널리 사용되는 균일 잡음(uniform noise)과 덩어리 잡음(blob noise)을 사용하였다. 덩어리 잡음은 먼저 잡음 덩어리의 중심을 균일 잡음 모델로부터 생성하고, 생성된 중심에서 가우시안 잡음을 생성하였다.

PFCM-R을 이용하기 위해서는 regularization 상수 β를 결정하여야 한다. 이 논문에서는 먼저 잡음이 없는 데이터에서 최소의 클러스터링 에러를 나타내는 β를 결정하고 이를 잡음이 첨가된 데이터에서도 사용하였다. 클러스터링 에러는 잘못 할당된 데이터 포인트의 개수를 전체 데이터 포인트의 개수로 나눈 값이며, 잡음 비율은 잡음 포인트의 개수를 전체 데이터 포인트의 개수로 나눈 값이다.

표 1. 균일 잡음을 이용한 경우의 클러스터링 결과
Table 1. Clustering results with uniform noise

잡음 비율	방법	D_2	D_3	D_4
0%	PFCM	0.0204	0.0398	0.0388
	PFCM-R	0.0211	0.0327	0.0318
50%	PFCM	0.0237	0.0808	0.0673
	PFCM-R	0.0236	0.0400	0.0356
100%	PFCM	0.0264	0.1311	0.0967
	PFCM-R	0.0264	0.0513	0.0396

표 1은 균일 잡음을 이용한 경우의 클러스터링 결과를 요약한 것이다. D_2의 경우 데이터가 단순하므로 PFCM과 PFCM-R이 거의 동일한 결과를 보이고 있다. 하지만 D_3와 D_4의 경우 잡음 비율이 증가함에 따라 PFCM-R이 PFCM보다 나은 성능을 보임을 알 수 있다. 그림 2는 D_3에서 균일 잡음 비율에 따른 에러 비율을 나타낸 그래프이다.

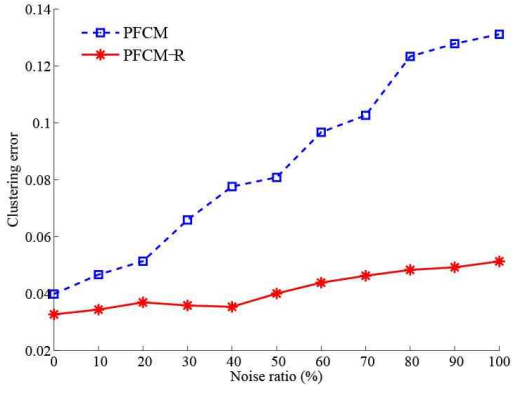


그림 2. D_3에서 균일 잡음 비율에 따른 클러스터링 에러
Fig. 2. Cluster error rates of D_3 with respect to uniform noise ratio

그림에서 보아 알 수 있듯이, PFCM-R은 잡음 비율이 증가함에도 불구하고 클러스터링 에러가 크게 증가하지 않는다. 이는 regularization이 해공간을 평탄화함으로써 국부 최적해에 빠지는 가능성을 줄여주기 때문이다. 두 방법의 에러 비율 차이는 잡음 비율에 비례하여 커지고 있다, 즉, 잡음 환경에서 PFCM-R이 보다 나은 성능을 보임을 알 수 있다. D_4에 대한 실험에서도 유사한 그래프를 얻을 수 있었다.

표 2. 덩어리 잡음을 이용한 경우의 클러스터링 결과
Table 2. Clustering results with blob noise

잡음 비율	방법	D_2	D_3	D_4
0%	PFCM	0.0174	0.0260	0.0293
	PFCM-R	0.0174	0.0259	0.0290
50%	PFCM	0.0257	0.0705	0.1644
	PFCM-R	0.0257	0.0705	0.1641
100%	PFCM	0.0477	0.0638	0.2685
	PFCM-R	0.0477	0.0638	0.2670

표 2는 덩어리 잡음을 이용한 경우 클러스터링 결과를 요약한 것이다. 균일 잡음의 경우에서와는 달리 PFCM-R은 PFCM과 거의 동일한 결과를 보여주고 있다. 이는 그림 3에서 보인 바와 같이, 덩어리 잡음에 의해 평탄화된 해공간에서도 국부 최적해에 빠지기 때문이다.

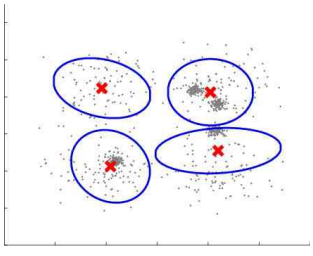


그림 3. D_4와 50% 덩어리 잡음에 대한 클러스터링 결과 예
Fig. 3. An example of clustering result of D_4 with 50% blob noise

비록 덩어리 잡음의 경우 PFCM-R이 PFCM과 거의 동일한 결과를 보여주었지만, 이는 β의 결정 방법에 따라 개선될 수 있다. 앞의 실험에서는 잡음 비율에 관계없이 동일한 β 값을 사용하였다. 하지만 잡음 비율에 따라 β 값을 수정하는 것이 보다 자연스럽다. 잡음 비율에 따라 β를 동적으로 결정하기 위해서는 먼저 잡음이 없는 데이터에서 β_0를 결정하고, 식 (15)의 첫 번째와 세 번째 항의 비율을 계산한다.

$$\xi = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik}^{m} + t_{ik}^{\eta})D_{ik}^{2}}{\beta_0 \sum_{i=1}^{c} u_{ik}^{m}} = \frac{T_{1,0}}{\beta_0 T_{3,0}} \quad \text{(24)}$$

잡음이 첨가된 데이터 집합이 주어지면, 먼저 PFCM을 이용하여 클러스터링을 수행하고, 식 (25)와 같이 ξ 값이 항상 일정한 값을 유지하도록 β를 결정한다.

$$\xi = \frac{T_{1,0}}{\beta_0 T_{3,0}} = \frac{T_1}{\beta T_3} \quad \text{(25a)}$$

$$\beta = \xi \frac{T_3}{T_1} \quad \text{(25b)}$$

그림 4는 β의 결정 방법에 따른 D_4의 클러스터링 에러를 표시한 것이다. 그림에서 보아 알 수 있듯이, ξ 값이 일정한

값을 유지하도록 하는 경우, 즉, 잡음 비율에 따라 β 값을 달리 결정하는 경우, β 값을 일정한 값으로 유지하는 경우에 비해 나은 결과를 얻을 수 있었다. 하지만 여전히 β_0를 결정하는 문제가 남아 있으며 이는 현재 연구 중에 있다.

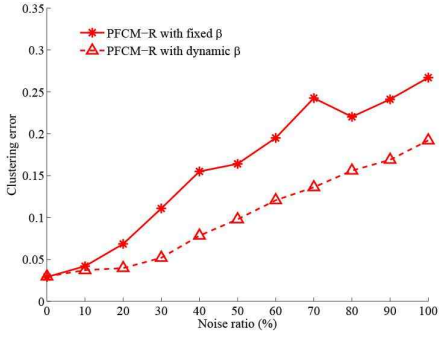


그림 4. β 결정 방법에 따른 D_4의 클러스터링 에러

Fig. 4. Clustering errors of D_4 with different methods for deciding β

V. 결론

PFCM은 FCM과 PCM의 장점으로 서로의 단점을 보완함으로써 이전의 퍼지 클러스터링 방법들에 비해 나은 결과를 보였다. 하지만 잡음 비율이 증가함에 따라 PFCM 역시 잡음의 영향을 받게 되며, 이 논문에서는 regularization을 이용하여 PFCM의 잡음 민감성을 줄인 PFCM-R을 제안하였다. 제안한 방법은 PFCM에 비해 잡음이 많은 경우 보다 정확하게 클러스터 구조를 찾아낼 수 있음을 실험 결과를 통해 확인할 수 있었다.

PFCM-R이 기존의 방법에 비해 나은 결과를 보여주기는 했지만, regularization 상수 β를 결정하는 문제는 regularization을 사용하는 다른 방법들에서와 마찬가지로 해결해야 할 과제로 남아 있다. 비록 이 논문에서 사용한 격자 탐색법이 좋은 결과를 보여주었지만, 이는 많은 시간을 요하는 단점이 있다. 또한 잡음 비율에 따라 β를 동적으로 설정하는 경우 보다 나은 결과를 얻을 수 있음을 확인하였지만, 이 경우에서도 역시 β_0를 효과적으로 결정하는 방법이 필요하다. 주어진 데이터의 특성에 따라 β 또는 β_0를 결정하는 방법은 현재 연구 중에 있다.

참고문헌

[1] L. A. Zadeh, “Fuzzy sets,” Information and Control, Vol. 8, No. 3, pp. 338 - 353, 1965.

[2] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, New York, Springer, 1981.

[3] R. Krishnapuram and J. Keller, “A possibilistic approach to clustering,” IEEE Transactions on Fuzzy Systems, Vol. 1, No. 2, pp. 98 - 110, 1993.

[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. Zhang and Y. W. Leung, “Improved possibilistic c-means clustering algorithms,” IEEE Transactions on Fuzzy Systems, Vol. 12, No. 2, pp. 209 - 217, 2004.

[6] N. R. Pal, K. Pal, and J. Bezdek, “A mixed c-means clustering model,” Proceedings of 6th IEEE International Conference on Fuzzy Systems, July 1997, pp. 11 - 21, 1997.

[7] A. Tikhonov, “On solving incorrectly posed problems and method of regularization,” Dokl. Acad. Nauk USSR, Vol. 151, pp. 501 - 504, 1963.

[8] R. Li and M. Mukaidono, “A maximum entropy approach to fuzzy clustering,” Proceedings of 4th IEEE International Conference on Fuzzy Systems, March 1995, pp. 2227 - 2232, 1995.

[9] S. Miyamoto and K. Umayahara, “Fuzzy clustering by quadratic regularization,” Proceedings of 4th IEEE International Conference on Fuzzy Systems, March 1998, pp. 1394 - 1399, 1998.

[10] D. Ozdemir and L. Akarun, “A fuzzy algorithm for color quantization of images,” Pattern Recognition, Vol. 35, No. 8, pp. 1785 - 1791, 2002.

[11] J. Yu and M. S. Yang, “A generalized fuzzy clustering regularization model with optimality tests and model complexity analysis,” IEEE Transactions on Fuzzy Systems, Vol. 15, No. 5, pp. 904 - 915, 2007.

[12] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," SIAM Review, Vol. 34, No. 4, pp. 561-580, 1992.

[13] D. K. Stando and M. Rudnicki, “Regularization parameter selection in discrete ill-posed problems - the use of the U-curve,” International Journal of Applied Mathematics and Computer Science, Vol. 17, No. 2, pp. 157-164, 2007.
[14] C. W. Hsu, C. C. Chang, and C. J. Lin, “A practical guide to support vector classification,” Technical Report, 2003, Department of Computer Science, National Taiwan University, Taipei, Taiwan.

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

남 궁 영 환

2003년 5월 : University of Southern California (이학석사)
2003년 8월~현재: Department of Computer and Information Science and Engineering, University of Florida (박사과정)
관심분야 : 인공지능, 데이터 마이닝, 패턴인식

김 성 훈

1996년 2월 : 연세대학교 본대학원 전자공학과 (공학박사)
1996년 3월 ~ 2006년 2월 : 영동대학교 컴퓨터공학과 부교수
2006년 3월~현재: 경북대학교 컴퓨터정보학부 조교수
관심분야 : 인공지능, 패턴인식, 지능형콘텐츠