클러스터 밀도에 무관한 향상된 클러스터링 기법

유병현 · 김완우 · 허경용

동의대학교 전자공학과

An Improved Clustering Method with Cluster Density Independence

Byeong-Hyeon Yoo · Wan-Woo Kim · Gyeongyong Heo

Dept. of Electronic Engineering, Dong-eui University

E-mail : youyooo@naver.com, wwkim614@naver.com, hgycap@deu.ac.kr

요 약

클러스터링은 대표적인 비교사 학습 방법의 하나로 균일한 특성을 가지는 데이터를 클러스터로 묶기 위해 사용된다. 하지만 클러스터링은 기본적으로 클러스터의 중심에서 데이터까지의 거리에 기반하고 있으므로 클러스터의 중심이 밀도가 높은 클러스터 쪽으로 쏠리는 현상이 발생한다. 이 논문에서는 클러스터의 중심을 가능한 멀리 떨어져 있도록 하는 항을 Fuzzy C-Means의 목적함수에 추가함으로써 클러스터 사이의 밀도 차이가 심한 데이터의 클러스터링 문제에서 정확한 결과를 얻을 수 있는 클러스터링 방법을 제안한다. 제안한 방법은 FCM에 비해 실제 클러스터 중심으로 수렴하는 경우가 더 많으며 수렴 속도 역시 FCM 보다 빠른 것을 실험 결과를 통해 확인할 수 있다.

ABSTRACT

Clustering is one of the most important unsupervised learning methods that clusters data into homogeneous groups. However, cluster centers tend leaning to high density clusters because clustering is based on the distances between data points and cluster centers. In this paper, a modified clustering method forcing cluster centers to be apart by introducing a center-scattering term in the Fuzzy C-Means objective function is introduced. The proposed method converges more to real centers with small number of iterations compared to the original one. All the strengths can be verified with experimental results.

키워드

클러스터링, FCM, 밀도 무관성

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, 퍼지 집합은 클러스터링 분야에 도입되었고 퍼지 클러스터링은 대표적인 클러스터링 기법 중 하나로 자리 잡았다. Bezdek[1]에 의해 일반화된 Fuzzy C-Means(FCM)는 퍼지 클러스터링의 대표적인 방법 중 하나이다. FCM은 간단하면서도 효과적인 클러스터링 방법이지만, 구해진 소속도가 직관적인 값과 일치하지 않는 경우가 있으며, 클러스터의 밀도가 서로 다른 경우 클러스터의 중심이 밀도가 높은 클러스터 쪽으로 치우치는 경향이 있다. 이 논문에서는 밀도 차이에 의해 클러스터의 중심이 치우치는 경우를 해결하기 위한 방법으로 클러스터 중심 사이의 거리를 나타내는 항을 목적함수에 포함시킨 새로운 클러스터링 기법을 제시한다.

Ⅱ. 본 론

퍼지 클러스터링은 제약 조건이 있는 최적화 문제(constrained optimization problem)로[2 3], 식 (1)의 목적 함수를 최적화하는 것으로 볼 수 있다.

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

이때 $1 < m < \infty$은 퍼지화 정도를 나타내는 상수로 일반적으로 2로 설정되며, n은 데이터 포

인트의 개수를, c는 클러스터의 개수를, v_i는 i번째 클러스터의 중심을, u_{ik}는 k번째 데이터 포인터가 i번째 클러스터에 소속되는 정도를, x_k는 k번째 데이터 포인터를 나타낸다. 이 논문에서는 유클리드 거리 척도를 사용하였다.

FCM의 밀도가 다른 경우 클러스터 중심이 치우치는 현상은 FCM이 클러스터 중심에서 데이터까지의 거리에 기반[4 5]하고 있기 때문에 나타난다. 이 같은 클러스터 중심이 쏠리는 현상을 방지하기 위해 이 논문에서는 클러스터의 중심이 가능한 멀리 떨어져 있도록 하는 항을 FCM의 목적함수에 추가하였다. 제안하는 변형된 FCM의 목적함수는 식 (2)와 같다.

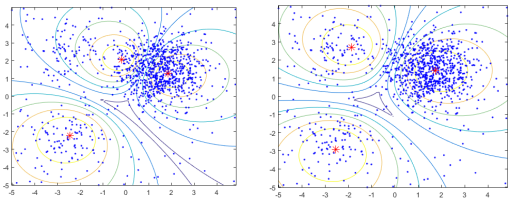
$$J=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^{m}\|v_i-x_k\|^2-\alpha\sum_{a=1}^{c}\sum_{b=1}^{c}\|v_a-v_b\|^2 \qquad (2)$$

이 때 α(> 0)는 중심이 떨어져 있는 정도를 목적함수에 반영하는 비율을 나타내는 상수이다. 식 (2)를 최적화시키는 u_{ik}와 v_i를 라그랑즈(Langrange) 방법을 사용하여 식 (3) 및 식 (4)와 같이 구할 수 있다.

$$u_{ik}=\frac{\frac{1}{\|v_i-x_k\|^2}}{\sum_{i=1}^{c}\frac{1}{\|v_i-x_k\|^2}} \qquad (3)$$

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}^2x_i-\alpha\sum_{b=1}^{c}u_b}{\sum_{k=1}^{n}u_{ik}^2-\alpha c} \qquad (4)$$

소속도를 구하는 식 (3)은 기존의 FCM과 같지만, 클러스터의 중심을 구하는 식 (4)에서 FCM과 차이가 있다. 그림 1-(a)는 샘플 데이터 집합에 FCM을 적용한 경우의 클러스터링 결과로, 클러스터의 중심이 밀도가 높은 클러스터 쪽으로 치우치고 있음을 알 수 있다. 반면 제안한 방법을 적용한 경우에는 클러스터의 밀도 차이에도 불구하고 정확한 클러스터 중심을 찾아내고 있음다.



(a) FCM (b) 제안하는 방법

그림 1. 밀도가 다른 클러스터에 대한 클러스터링 결과

제안하는 방법에서는 α에 따라 클러스터 중심 사이 거리가 목적함수에 반영되는 정도가 달라지며 이는 데이터의 분포에 따라 달리 설정되어야 한다. 그림 1의 샘플 데이터 집합에 대해 α를 0.2에서 8까지 0.2 간격으로 변화시키면서 실제 샘플 데이터 생성에 사용된 클러스터 중심과 클러스터링 결과로 얻어진 클러스터 중심 사이의 거리를 오류값으로 계산하여 나타낸 것이 그림 2이다. 그림 2에서 알 수 있듯이 제안하는 방법은 FCM에 비해 평균 오류가 적었으며 특히 샘플 데이터의 경우 α가 5.6일 때 최소 평균 오류 2.39를 가져 FCM에 비해 오류가 31.84% 줄어들었다. 또한 수렴 속도에서도 평균 반복횟수 72.22로 9.73% 감소하였다.

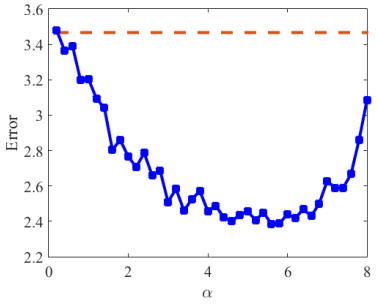


그림 2. 제안한 방법의 α에 따른 오류
(점선은 FCM의 경우 오류)

III. 결 론

이 논문에서는 클러스터의 밀도 차이로 인해 클러스터 중심이 왜곡되는 현상을 해결하기 위해 클러스터 중심 사이의 거리를 목적함수에 추가한 새로운 클러스터링 방법을 제안하였다. 실험결과에서 제시한 바와 같이 제안하는 방법은 FCM에 비해 실제 클러스터 중심에 수렴하는 확률이 높고 수렴 속도 역시 FCM에 비해 빠른 것을 알 수 있다.

참고문헌

[1] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, New York, Springer, 1981.
[2] Sadaaki Miyamoto, Fuzzy Clustering - Basic Ideas and Overview, Springer Handbook of Computational Intelligence pp239-248, 2015
[3] Janmenjoy Nayak, Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014, Volume 32 of the series Smart Innovation, Systems and Technologies pp 133-149, 2014
[4] Zarita Zainuddin, An effective fuzzy C-means algorithm based on symmetry similarity approach, volume 35 of Applied Soft Computing pp 433-448, 2015
[5] Basel Abu-Jamous, Fuzzy Clustering, 2015