새로운 퍼지 군집화 알고리즘

김재영*, 박동철*, 한지호*, HUYNH THI THANH THUY*, 송영수**
명지대학교*, 팬텍계열**

A New Fuzzy Clustering Algorithm

Jae-Young Kim*, Dong-Chul Park*, Ji-Ho Han*, HUYNH THI THANH THUY*, Young-Soo Song**
Myong Ji University*, PANTECH Group**

Abstract - 본 논문은 데이터의 군집화를 효율적으로 수행하기 위하여 새로운 군집화 알고리즘을 제안한다. 제안되는 군집화 알고리즘은 Fuzzy C-Means (FCM)에 기반을 두는데, FCM 알고리즘은 모든 데이터에 대한 거리에 기본을 둔 멤버쉽을 기초로 하기 때문에 잡음에 약한 제약을 지니고 있었다. 이를 개선하기 위하여, 제안되었던 PCM(Probabilistic C-Means), FPCM(Fuzzy PCM), PFCM(Probabilistic FCM) 등 여러가지 알고리즘이 제안되었다. 그러나 이들 알고리즘들은 초기 파라미터값 설정과 과다한 계산양에 따른 문제가 증가하였으며, 또한 잡음에 어느 정도 민감한 문제점을 지니고 있었다. 이 논문에서는 잡음에 대해 효과적으로 대응할 수 있는 새로운 군집화 알고리즘을 제안하고, 전통적인 군집화를 위한 Iris 데이터에 대한 실험을 통하여 효용성을 확인하였다.

1. 서 론

퍼지 군집화 방법은 오랫동안 많은 연구가 진행되고 있다. 이 방법은 남자나 여자, 0아니면 1과 같이 확실하게 구분 짓기 보다는 그 중간 값들을 허용하여 모든 데이터에 대한 군집화를 판단할 수 있는 이론을 토대로 발전하였다. 그러므로 퍼지 군집화 방법을 이용할 때는 하나의 군집에 대해서만 소속을 정하지 않고 전체 군집에 대한 소속정도를 계산한다. 이와 같은 방법을 이용한 퍼지 군집화 방법에 대한 관심과 그 효용성을 여러 가지 응용을 통해 퍼지 군집화에 대한 성과를 가져왔다.

퍼지 군집화 방법 중 가장 대표적인 알고리즘이 Fuzzy C-Means (FCM) 알고리즘[1]이다. FCM 알고리즘은 멤버쉽을 이용해 각 군집을 학습시키며, 이 멤버쉽들은 하나의 데이터에 대한 모든 군집의 멤버쉽의 합을 1로 제한함으로 구해진다. 그러나 FCM 알고리즘은 잡음에 대해 효과적으로 학습할 수 없으므로 Outlier와 같은 잡음이 많은 데이터에 대해서는 사용이 제한된다.

FCM으로부터 발전된 PCM(Probabilistic C-Means)[2], FPCM(Fuzzy PCM)[3], PFCM(Probabilistic FCM)[4]등의 알고리즘들은 FCM 알고리즘의 잡음에 민감한 부분을 제거하기 위하여 개발된 알고리즘들이다. 그러나 잡음에 대한 성능의 향상과 함께,계산량도 상대적으로 많이 증가하게 되었다. 본 논문에서는 FCM 알고리즘의 잡음문제를 새롭게 접근해보고, 새로운 멤버쉽 함수를 정의함으로써 잡음에 대한 문제를 해결하고자 한다. 또한 Iris 데이터에 대한 실험을 통해 그 효용성을 확인하였다.

이 논문은 새롭게 제안된 알고리즘을 2장에서 소개하고, 3장에서는 기존의 잘 알려진 Iris 데이터를 이용하여 실험을 통해제안된 알고리즘의 성능을 평가하였으며, 4장에서는 본 논문의 결론을 기술하였다.

2. 제안된 Intuitive Fuzzy C-Means Algorithm

PCM, FPCM, PFCM 알고리즘들은 FCM 알고리즘의 잡음 문제를 제거하기 위하여 제안되었으나, 잡음 문제 이외의 더 많은 문제점들을 만들고 있다. 그러므로 본 논문에서는 이러한 잡음 문제점을 보다 효과적으로 제거하기 위해 새로운 알고리즘인 Intuitive Fuzzy C-Means (IFCM) 알고리즘을 제안한다.

IFCM 알고리즘은 식(1)와 같이 새로운 목적함수를 정의 하였다.

$$\min(I,V)\left\{J_m(I,V;X)=\sum_{k=1}^{n}\sum_{i=1}^{c}(I_{ik})\|X_k-V_i\|^2\right\} \tag{1}$$

기본적으로 FCM 알고리즘의 목적함수와 동일하나 멤버쉽 대신 직관적인 멤버쉽(Intuitive Membership : I_{ik})를 사용한다. I_{ik}는 기존의 멤버쉽을 이용하여 식(2)와 같이 계산된다.

$$I_{ik}=\sum_{l=1}^{c}(u_{ik})^{\eta}|u_{ik}-u_{lk}| \tag{2}$$
$$1\le i\le c, 1\le k\le n$$

I_{ik}는 식(2)를 통해 잡음(Outlier)에 대한 높은 멤버쉽 값을 부여했던 FCM 알고리즘을 개선하기 위해 고안되었으며, 각 군집에 대해 서로 배타적인 성격을 가지게 되어 직관적인 분류 효과를 가져 올 수 있다. 또한 η을 이용하여 멤버쉽의 가중치 정도를 선택할 수 있다.

$$u_{ik}=\frac{1}{\sum_{j=1}^{c}\left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{2}{m-1}}} \tag{3}$$
$$1\le i\le c, 1\le k\le n$$

멤버쉽 함수는 기존의 FCM 알고리즘의 멤버쉽을 사용하며 식(3)과 같이 사용한다. IFCM 알고리즘의 대표값을 학습하기 위해 식(4)를 사용하며, 식(4)는 아래와 같이 유도 된다.

$$J_i(I,V;X)=\sum_{k=1}^{n}(I_{ik})(X_k-V_i)^2,\ 1\le i\le c$$
$$\frac{\partial J_k}{\partial V_i}=\sum_{k=1}^{n}(I_{ik})(-2)(X_k-V_i)=0$$
$$\sum_{k=1}^{n}(I_{ik})X_k-\sum_{k=1}^{n}(I_{ik})V_i=0$$

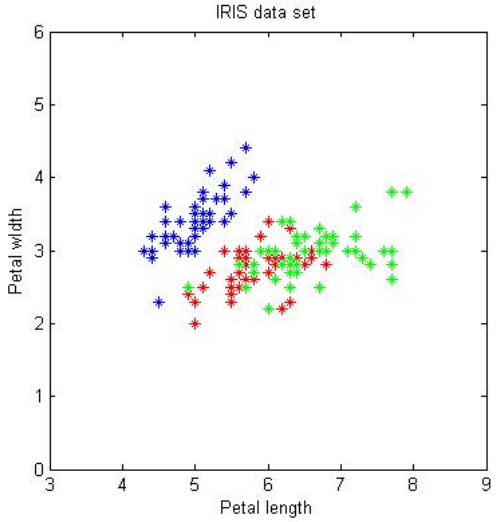
$$\therefore\ V_i=\frac{\sum_{k=1}^{n}(I_{ik})X_k}{\sum_{k=1}^{n}(I_{ik})},\ 1\le i\le c \tag{4}$$

IFCM 알고리즘은 식(3)과 (4)를 이용하여 반복적으로 각 군집의 대푯값들을 학습한다. 이 알고리즘은 식(2)와 같이 멤버쉽에 대한 제약조건은 여전히 존재한다. 그러나 I_{ik}가 $2\le c\le n$ 라면 다음의 제약조건을 가진다.

$$0\le I_{ik}\le c-1,\quad \forall i,k \tag{5}$$

<표 1>은 IFCM알고리즘의 pseudo-code 이다. IFCM 알고리즘은 PCM이나 FPCM, PFCM 알고리즘과 같이 초기화에 민감하지 않고, 대용량 데이터에 이용 가능하며, 파라미터 설정 값들이 존재 하지 않으므로 효과적으로 데이터를학습하고 결과 값을 도출 할 수 있다. 이러한 IFCM 알고리즘의 장점은 3장의 실험을 수행하여, 각 알고리즘의과의 성능 비교를 통해 확인하였다.

본 연구는 한국과학재단 특정기초연구
(R01-2007-000-20330-0) 지원 에 의한 것임.



<그림 1> 꽃잎의 폭과 길이에 대한 Iris 데이타

3. 실험 및 결과

데이터 군집화 분야에서 실험을 위하여 많이 쓰이고 있는 Iris 데이터[5]를 이용하여 알고리즘의 유용성을 확인하였다. 이 데이터는 Setosa, Versicolor, Virginica의 3개의 군집으로 되어 있으며, 각 군집은 50개의 데이터로 구성되어 있다. 또한 각 데이터는 꽃잎의 폭과 길이, 꽃받침의 폭과 길이의 4차원으로 구성되어 있다. 그림 1은 Iris 데이터의 꽃잎의 폭과 길이에 대해서 2차원으로 나타낸 그림이다. 이 그림에서 보듯이 2차원 상에서는 한 개의 군집은 다른 두 개의 군집과과 거의 분리가 되지만 나머지 2개의 군집은 서로 겹쳐져 섞여 있는 것을 볼 수 있다.

실험에서도 PCM 알고리즘은 FCM 알고리즘을 이용해 학습한 후 다시 PCM 알고리즘을 이용하여 다시 학습하여 초기값 설정하였다. 또한 FCM, PCM, FPCM, PFCM 알고리즘의 파라미터 값들을 변화를 주어 실험하였고 표 2에서 보는 봐와 같이 결과를 얻을 수 있었다.

<표 2>는 Iris 데이터를 학습한 후의 분류 오류값과 정확도에 대해 나타내었다. 각각의 알고리즘은 여러 차례의 실험을 통해 파라미터들에 대해 변화를 주어 실험하였으며 최고의 결과값을 나타낸 것이다. 실험에서 보는 바와 같이 대부분의 알고리즘이 분류정확도에서 90%의 성능을 보여 주었지만 IFCM 알고리즘은 다른 알고리즘에 비해 상당히 잘 분류해 내며 92.67%까지의 정확도를 보여준다. 또한 다른 알고리즘은 Virginica에 대해 분류정도가 상당히 낮았으나 IFCM 알고리즘은 상대적으로 높은 성능을 보여준다.

4. 결 론

퍼지 군집화 알고리즘은 오래된 역사를 가지고 있다. 개발된 여러 가지 알고리즘은 FCM 알고리즘의 잡음 문제를 해결하기 위하여 PCM, FPCM, PFCM과 같은 알고리즘들로 발전하였다. 하지만 이 알고리즘들은 완벽하게 잡음에 대처하지 못하였으며, 초기화 문제, 계산양의 증가, 파라미터들의 증가로 실제적인 군집화 과정에 이용하기가 복잡해 졌다.

이를 해결하기 위하여 본 논문에서는 IFCM 알고리즘을 새롭게 제안하였다. IFCM 알고리즘은 FCM 알고리즘의 멤버쉽을 대체하는 직관적인 멤버쉽(Intuitive Membership)을 고안하여 효과적으로 군집들을 학습할 수 있었다. 또한 잡음이나 다른 군집에 속한 데이터들에 대해 상대적으로 작은 멤버쉽을 부여하기 때문에 각각의 군집을 분류하기 위해 효과적이었다. IFCM 알고리즘은 FCM 알고리즘의 잡음 문제를 해결하였으며, PCM 알고리즘의 초기값 문제, FPCM 알고리즘의 대용량 데이터 처리 문제, PFCM 알고리즘의 파라미터 설정 문제들에 대해 효과적으로 해결할 수 있었다. 앞으로의 연구에서는 다양한 데이터에 대한 더 많은 실험을 통해 잡음에 대한 성능 확인을 시도하려 한다.

<표 1> IFCM의 pseudo-code

```
Algorithm IFCM
  Procedure main()
    Read c, m, ε
    [c: initialize cluster,
     m: weighting exponent(m∈1,...∞)]
    error = 0
    While (error > ε)
      While (input file is not empty)
        Read the data x
        [Update IFCM Membership]
          u_ik = 1 / Σ_{j=1}^{c} (D_ik / D_jk)^{2/(m-1)}
          1 ≤ i ≤ c, 1 ≤ k ≤ n
        [Calculate IFCM Intuitive Membership]
          I_ik = Σ_{l=1}^{c} (u_ik)^η |u_ik − u_lk|
          1 ≤ i ≤ c, 1 ≤ k ≤ n
        [Update IFCM Center Mean]
          V_i = Σ_{k=1}^{n} (I_ik) X_k / Σ_{k=1}^{n} (I_ik)   , 1 ≤ i ≤ c
          e := v(n+1) − v(n)
      End While
        error = e
    End While
  End main()
End
```

<표 2> IRIS 데이터의 실험 결과

Algorithm	Setosa	Versicolor	Virginica	분류 오류	정확도	기타
FCM	50	47	38	15/150	90.00%	m=3
PCM	50	47	39	14/150	90.67%	m=3
FPCM	50	47	38	15/150	90.00%	m=3, η=3
PFCM	50	45	41	14/150	90.67%	m=2, η=2, a=1, b=1
IFCM	50	45	44	11/150	92.67%	m=3, η=2

[참 고 문 헌]

[1] J. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*, New York: Plenum, 1981.

[2] R. Krishnapuram and J. Keller, " A possibilistic approach to clustering," *IEEE Trans. on Fuzzy Syst.* V.1, No. 2, pp. 98-110, 1993.

[3] R. Krishnapuram, H. Frigui, and O. Nasroui, " Fuzzy and possibilistic shell clustering algorithm and their application to boundary detection and surface approximation," *IEEE Trans. on Fuzzy Syst.* V.3, No. 1, pp. 29-60, 1995.

[4] N. Pal, K. Pal, and J. Bezdek," A possibilistic fuzzy c-means algorithm," *IEEE. Tr. Fuzzy Syst.* V.13, No. 4, pp. 517-530, 2005.

[5] E. Anderson, "The irises of the GASPE peninsula", *Bulletine of American Iris Society*, V. 59, pp. 2 - 5, 1935,