퍼지값을 갖는 데이터에 대한 퍼지 클러스터링

이 건 명
충북대학교 컴퓨터과학과

Fuzzy Clustering for Fuzzy Data[1]

Keon-Myung Lee
Dept of Computer Science, ChungBuk National Univ

요 약

클러스터링은 데이터의 특성 추출, 데이터의 압축 등을 목적으로 동일 클러스터에 속하는 데이터간에는 유사성이 크도록 하면서 다른 클러스터에 속하는 데이터간에는 유사성이 작도록 데이터를 군집화하는 것이다. 일상에서 발생하는 많은 데이터에는 관측 오류, 불확실성, 주관적인 판정 등으로 인해서 데이터의 속성값이 정확한 값으로 주어지지 않은 경우가 있다. 본 논문에서는 분명한 값뿐만 아니라 퍼지값도 포함한 데이터들에 대해서 퍼지 클러스터링하는 방법을 제안한다

1. 서론

클러스터링(clustering)은 데이터의 특성 추출, 데이터의 압축 등을 목적으로 동일 클러스터에 속하는 데이터간에는 유사성이 크도록 하면서 다른 클러스터에 속하는 데이터간에는 유사성이 작도록 데이터를 군집화하는 것이다. 일상에서 발생하는 데이터에는 관측 오류, 불확실성, 주관적인 판정 등으로 인해서 데이터 자체에 애매함이 포함되는 경우가 많다. 이러한 애매함을 표현하는 효과적인 방법의 하나가 퍼지 집합을 사용하는 방법이다[5]. 애매한 값이 포함된 데이터 표현에 퍼지 집합을 사용하는 것은 데이터 표현에 융통성을 주지만, 이러한 데이터들을 대상으로 하는 클러스터링 방법이 아직 없기 때문에 데이터 처리가 곤란하다. 따라서 본 논문에서는 분명한 값뿐만 아니라 퍼지 집합으로 표현된 애매한 값을 포함한 데이터에 대해서도 클러스터링할 수 있는 방법을 제안한다.

본 논문은 다음과 같이 구성된다. 2 절에서는 퍼지값을 포함한 데이터의 형태 및 표현방법에 대해서 설명하고, 3 절에서는 분명한 값, 구간값 및 퍼지값이 혼합된 데이터간의 거리척도로서 비유사성(dissimilarity) 척도를 제안하고, 클러스터중심의 표현방법을 소개한다. 4 절에서는 퍼지값을 포함한 이러한 데이터에 대한 퍼지 클러스터링 방법을 소개하고, 5 절에서는 제안된 방법에 대한 실험결과를 살펴본 다음, 끝으로 결론을 맺는다

2. 데이터의 표현

일반적으로 데이터 D는 여러 개의 속성값($A_1, A_2, \ldots, A_n$)으로 구성된다. 데이터를 구성하는 속성값은 분명한(crisp) 값뿐만 아니라 애매한 값(즉, 구간값, 퍼지값) 등을 갖을 수 있다. 속성은 특성에 따라 연속적인 값을 갖는 것과 이산적인 값을 갖는 것으로 나눌 수 있다. 연속적인 값을 갖는 속성은 분명한 값, 구간 값, 퍼지숫자(fuzzy number) 등의 값을 갖을 수 있고, 이산적인 값을 갖는 속성은 분명한 값, 이산영역에 대한 퍼지 집합 등의 값을 갖을 수 있다. 분명한 값, 구간값, 퍼지 집합, 퍼지숫자를 동일 플랫폼에서 처리하기 위해서는 이들 값에 대한 보편(universal) 표현법을 사용하는 것이 편리하다. 따라서 여기에서는 연속영역(continuous domain)의 분명한 값, 구간값, 퍼지숫자를 나타낼 때는 다음과 같이 사다리꼴

[1] 본 연구는 98 년 '대학기초연구지원사업'지원을 받아 수행된 것임

(trapezoidal) 퍼지숫자 $Trap(\alpha, \beta, \gamma, \delta)$를 사용한다. 연속영역에 대한 애매한 값을 나타내는 퍼지값은 사다리꼴 퍼지숫자로 근사하여 표현한다고 가정한다.

$$Trap(\alpha,\beta,\gamma,\delta)=\begin{cases}0 & \text{if } x<\alpha\\ (x-\alpha)/(\beta-\alpha) & \text{if } \alpha\le x<\beta\\ 1 & \text{if } \beta\le x\le\gamma\\ (x-\delta)/(\gamma-\delta) & \text{if } \gamma<x\le\delta\\ 0 & \text{if } x>\delta\end{cases}$$

분명한 값 $a = Trap(a, a, a, a)$

구간값 $[a, b] = Trap(a, a, b, b)$

퍼지숫자 $F = Trap(a, b, c, d)$

분명한 값과 퍼지집합을 갖을 수 있는 이산영역(discrete domain)의 값은 다음과 같이 모두 퍼지집합으로 표현한다. 분명한 값은 소속정도가 1인 퍼지집합으로 표현할 수 있다.

분명한 값 $a = \{1/a\}$

퍼지집합 $A = \{(x, \mu_A(x)) \mid x \in U\}$

3. 거리 척도 및 클러스터 중심표현

데이터에 대한 클러스터링을 위해서는 데이터간의 거리 및 데이터와 클러스터 중심간의 거리를 측정할 수 있는 거리척도(distance measure)와, 데이터의 클러스터에 대한 소속정도에 기반하여 클러스터의 중심을 표현하는 방법이 고안되어야 한다. 여기에서는 거리척도로서 Gowda[4] 등에 의해 제안된 비유사성(dissimilarity)척도에 바탕을 둔 퍼지값을 포함한 데이터에도 적용할 수 있는 비유사성 척도를 제안한다. 임의의 두 데이터 $A = (A_1, A_2, \ldots, A_n)$, $B = (B_1, B_2, \ldots, B_n)$에 대해 제안된 비유사성 척도 $D(A, B)$는 다음과 같다. 여기서 A_k, B_k는 각각 k번째 속성값을 나타내고 s_k는 k번째 속성의 중요도를 나타낸다. $D_p(A_k, B_k)$는 위치(position)에 의한 거리, $D_s(A_k, B_k)$는 폭(span)

$$D(A,B)=\sum_{k=1}^{d} s_k D(A_k,B_k)$$

에 의한 거리, $D_c(A_k, B_k)$는 내용(content)에 의한 거리를 나타낸다.

$$D(A_k,B_k)=s_pD_p(A_k,B_k)+s_sD_s(A_k,B_k)+s_cD_c(A_k,B_k)$$

다음은 연속영역 속성에 대한 비유사성 척도를 나타낸다. 연속영역의 속성값은 보편 표현법에 의해 모두 사다리꼴 퍼지숫자로 나타낼 수 있다. 아래에서 A_k^α, B_k^α는 퍼지숫자에 대한 α-cut[5]을 나타낸다.

- 위치에 의한 거리 $D_p(A^\alpha, B^\alpha)$

$$D_p(A_k,B_k)=\int_0^1 \frac{|\text{median of } A_k^\alpha - \text{median of } B_k^\alpha|}{\text{length of maximum interval}}d\alpha$$

- 폭에 의한 거리 $D_s(A^\alpha, B^\alpha)$

$$D_s(A_k,B_k)=\int_0^1 \frac{|\text{length of } A_k^\alpha - \text{length of } B_k^\alpha|}{\text{span length of } A_k^\alpha \text{ and } B_k^\alpha}d\alpha$$

- 내용에 의한 거리 $D_c(A^\alpha, B^\alpha)$

$$D_c(A_k,B_k)=\int_0^1 \frac{|\text{length of } A_k^\alpha + \text{length of } B_k^\alpha - 2*\text{length of } A_k^\alpha \cap B_k^\alpha|}{\text{span length of } A_k^\alpha \text{ and } B_k^\alpha}d\alpha$$

다음은 이산영역 속성에 대한 비유사성 척도이다.

- 위치에 의한 거리 $D_p(A^\alpha, B^\alpha)$

$$D_p(A_k,B_k)=\frac{\left|\frac{\sum_i \mu_{A_k}x_k}{\sum_i \mu_{A_k}} - \frac{\sum_i \mu_{B_k}x_k}{\sum_i \mu_{B_k}}\right|}{\text{length of maximum interval}}$$

- 폭에 의한 거리 $D_s(A^\alpha, B^\alpha)$

$$D_s(A_k,B_k)=\frac{\left|\sum_i \mu_{A_k}(x_i)-\sum_i \mu_{B_k}(x_i)\right|}{\sum_i \mu_{A_k\cup B_k}(x_i)}$$

- 내용에 의한 거리 $D_c(A^\alpha, B^\alpha)$

$$D_c(A_k,B_k)=\frac{\sum_i \left|\mu_{A_k}(x_i)-\mu_{B_k}(x_i)\right|}{\sum_i \mu_{A_k\cup B_k}(x_i)}$$

클러스터에 대한 중심은 다음과 같이 계산한다. 여기에서 w_{ij}는 클러스터 C_j에 대한 데이터 X_i의 소속정도를 나타낸다. 연속영역인 k번째 속성에 대한 클러스터 중심의 값 $Trap(\alpha_k, \beta_k, \gamma_k, \delta_k)$은 다음과 같이 계산된다. 데이터 X_i에 대한 k번째 속성이 $x_{ik} = Trap(\alpha_{ik}, \beta_{ik}, \gamma_{ik}, \delta_{ik})$와 같이 표현된다고 가정한다.

$$\alpha_k=\frac{\sum_i (w_{ij})^m \alpha_{ik}}{\sum_i (w_{ij})^m} \quad \beta_k=\frac{\sum_i (w_{ij})^m \beta_{ik}}{\sum_i (w_{ij})^m} \quad \gamma_k=\frac{\sum_i (w_{ij})^m \gamma_{ik}}{\sum_i (w_{ij})^m} \quad \delta_k=\frac{\sum_i (w_{ij})^m \delta_{ik}}{\sum_i (w_{ij})^m}$$

이산영역의 속성인 경우에는 클러스터 중심값이 다음과 같이 퍼지집합 $\{(x_p, \mu_{C_{jk}}(x_p)) \mid x_p \in U\}$으로 표현된다.

$$\mu_{C_{jk}}(x_p)=\frac{\sum_i (w_{ij})^m \mu_{A_{ik}}(x_p)}{\sum_i (w_{ij})^m}, \quad x_p \in U,\ m \in (1,\infty)$$

4. 퍼지 데이터에 대한 퍼지 클러스터링

퍼지값을 포함한 데이터에 대한 클러스터링을 위해서 퍼지 c-means 알고리즘[5]에 기반한 다음과 같은 퍼지 클러스터링을 알고리즘을 사용한다

1 임의의 C 개의 초기 클러스터 중심 및 m 값을 선택한다.

2 **Repeat**

2.1 다음의 식을 통해서 각 데이터의 클러스터들에 대한 소속정도를 계산한다

$$w_{ij} = \begin{cases} 1 / \sum_{q=1}^{C} \left(\frac{\|X_i - Z_j\|}{\|X_i - Z_q\|} \right)^{2/(m-1)}, & q \neq i \\ 1, & Z_j = X_i \end{cases}$$

여기에서 $\|X_i - Z_j\|$는 3 절에서 제안한 비유사성 척도 $D(X_i, Z_j)$이다

2.2 3 절에서 제안한 클러스터 중심 계산방법을 사용하여 각 속성에 대한 클러스터 중심 $Trap(\alpha_k, \beta_k, \gamma_k, \delta_k)$을 계산한다

Until (수렴하거나 최대 수행횟수 초과할 때)

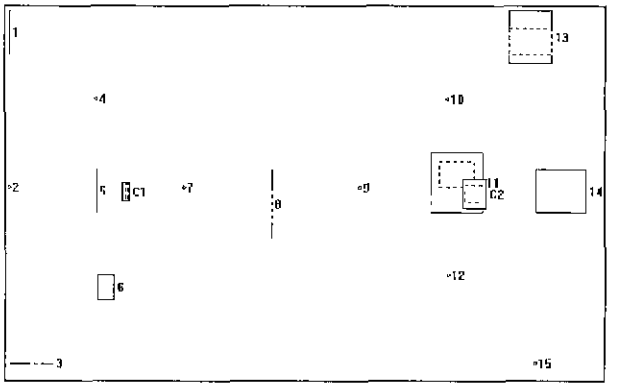
5. 실험

제안된 방법의 유용성을 보이기 위하여 몇가지 실험을 하였다 (그림 1)은 이중 하나의 실험 데이터 및 클러스터링 결과 찾은 클러스터 중심을 보인 것이다 실험에서 사용된 데이터는 두개의 연속영역 속성으로 구성된 것으로, 그림에서 점은 분명한 값을 나타내고, 직선이나 실선의 사각형만으로 구성된 것은 구간값이 포함된 데이터를 나타내고, 점선이나 점선의 사각형이 포함된 것이 퍼지숫자를 포함한 것을 나타낸다. 숫자는 데이터의 번호를 나타내고, C1 과 C2는 각각 클러스터의 중심을 나타낸다 실험에서는 비유사성 척도에서 위치에 의한 거리의 가중치를 3 로, 폭 및 내용에 의한 거리의 가중치는 1 로 하였고, m 값은 1 5로 하였다 (표 1)은 (그림 1)의 데이터에 대한 퍼지 클러스티링 결과이다 실험을 통해서 나온 결과가 직관과 부합됨을 알 수 있다

6. 결론

본 연구에서는 퍼지값을 포함한 데이터에 대해서 클러스터링 하는 방법을 연구하였다 여기에서는 거리척도로서 연속영역 및 이산영역에 대한 비유사성 척도를 제안하고, 퍼지값이 포함된 데이터에 대한 퍼지 클러스터링에서 클러스터 중심을 표현하는 방법을 제안하고, 이에 기반한 퍼지 클러스터링 알고리즘을 소개하였다 실험을 통해서 제안된 방법이 의미있는 결과를 도출할 수 있음을 보였다. 제안된 방법은 실세계에서 많이 발생하는 퍼지값을 포함한 데이터에 대한 클러스터링 분야에서 유용하게 활용될 수 있을 것으로 기대된다

(그림 1) 클러스티링 데이터



(표 1) 퍼지 클러스터링 결과

	1	2	3	4	5	6	7
C1	0 995	0 994	0 986	0 999	0 997	0.954	0 979
C2	0.005	0 006	0 014	0.001	0 003	0 046	0 021
8	9	10	11	12	13	14	15
0 706	0 290	0 271	0 040	0 264	0.052	0.007	0 027
0.294	0 710	0 729	0 960	0 736	0.948	0 993	0 973

참고 문헌

[1] Y El-Sonbaty, M A Ismail, Fuzzy clustering for symbolic data, *IEEE Trans. on Fuzzy Systems*, Vol 6. No 2, pp 195-204, May, 1998

[2] K Chidanada Gowda, E Diday. Symbolic clustering using a similarity measure, *IEEE Trans on System, Man, and Cybernetics*, Vol 22, No 2, pp 365-378, 1992

[3] K Chidanada Gowda, T V Ravi, Divisive clustering of symbolic object using the concepts of both similarity and dissimilarity, *Pattern Recognition*, Vol 28, No 8, pp.1277-1282, 1995

[4] K Chidanada Gowda, E Diday, Symbolic Clustering using a new dissimilarity measure, *Pattern Recognition*. Vol 24. No 6, pp.567-578, 1991

[5] Li-Xin Wang, *A Course in Fuzzy Systems and Control*, Prentice-Hall International, Inc., 424p. 1997.