퍼지 및 지능시스템학회 논문지 2003, Vol. 13, No. 6, pp. 661-666

범주형 데이터의 분류를 위한 퍼지 군집화 기법

A Fuzzy Clustering Algorithm for Clustering Categorical Data

김대원 · 이광형
Dae-Won Kim and Kwang H. Lee

한국과학기술원 전산학과

요 약

본 논문에서는 범주형 데이터의 분류를 위한 새로운 기법을 제시한다. 기존의 대표적인 퍼지 군집화 방법인 k-modes 알고리즘과 fuzzy k-modes 알고리즘은 군집의 중심을 단일 값으로 표현하고, 군집에 속하는 데이터의 빈도 수에 기반한 중심 갱신 기법을 사용하였다. 이와 같은 기존의 방법들은 분류의 경계가 모호한 데이터를 군집화할 경우, 알고리즘의 각 단계에서 발생하는 분류의 에러를 보정하지 못해 최종적으로 지역해에 빠지는 단점이 있다. 이를 극복하기 위해 본 논문에서는 군집 중심을 퍼지 집합을 이용하여 정의한다. 퍼지 군집 중심은 주어진 데이터와 군집간의 거리 관계를 퍼지 값을 이용해 표현하며, 각 군집의 중심은 데이터의 소속 정도 값을 이용해 갱신된다. 이와 같은 퍼지 중심 표현기법을 도입하여 범주형 데이터의 분류 시에 보다 세밀한 결정을 내림으로써, 인접한 군집들의 경계에서 발생하는 불확실성을 최소화한다. 기존의 대표적인 방법들과의 비교실험을 수행함으로써 제안한 방법의 성능을 검증하였다.

Abstract

In this paper, the conventional k-modes and fuzzy k-modes algorithms for clustering categorical data is extended by representing the clusters of categorical data with fuzzy centroids instead of the hard-type centroids used in the original algorithm. The hard-type centroids of the traditional algorithms had difficulties in dealing with ambiguous boundary data, which might be misclassified and lead to thelocal optima. Use of fuzzy centroids makes it possible to fully exploit the power of fuzzy sets in representing the uncertainty in the classification of categorical data. The distance measure between data and fuzzy centroids is more precise and effective than those of the k-modes and fuzzy k-modes. To test the proposed approach, the proposed algorithm and two conventional algorithms were used to cluster three categorical data sets. The proposed method was found to give markedly better clustering results.

Key words : 퍼지 클러스터링, 범주형 데이터, 퍼지 중심

1. 서 론

군집화 기법은 패턴인식, 영상처리 및 데이터마이닝 분야에서 매우 유용한 기법으로 사용되어왔다. 최근에는 데이터마이닝 연구에서 다양한 형태의 수치 데이터 및 범주형 데이터로 이루어진 데이터 집합을 처리하기 위해서 많은 연구가 이루어지고 있다. 비록 Gower의 유사도 척도를 이용한 계층적 군집화 기법을 비롯한 다양한 연구가 제안되어져 왔으나 [1-5], 이러한 기법들은 대규모 범주형 데이터를 포함한 데이터 집합에서는 효율성이 저하된다 [7].

이를 해결하기 위해서 k-means 알고리즘 형태의 기법이 제안되었다 [8-9]. Huang은 표준 k-means 알고리즘을 확장하여 새로운 유사도 척도와 빈도 수에 기반을 둔 k-modes 알고리즘을 제안하였다. 또한 k-modes 알고리즘을 Bezdek의 fuzzy c-means 알고리즘의 형태로 일반화한 fuzzy k-modes 알고리즘도 제안하였는데, 실제 데이터 집합에 응용함으로써 그 우월성을 제시하였다.

하지만 대부분의 퍼지 군집화 알고리즘에서, 군집의 중심 값은 하나의 스칼라 값으로 표현된다. 이러한 표현 형태는 불확실성과 모호함이 존재하는 공간에서 그 한계점을 지닌다. 따라서 본 논문에서는 퍼지 중심 값(fuzzy centroid) 표현을 가지는 새로운 군집화 기법을 제안한다. 이러한 퍼지 중심 값 표현은 불확실성을 해소하기 위해서 퍼지 집합 이론을 최대로 활용할 수 있으며, 지역해(local optima)에 빠지는 경향성을 감소시킬 수 있다.

접수일자 : 2003년 8월 27일
완료일자 : 2003년 11월 25일
본 논문은 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

2. Fuzzy k-modes 알고리즘

Fuzzy k-modes 알고리즘을 기술하기에 앞서 다음과 같이 입력 데이터 형태를 정의한다. $X=\{X_1,\ldots,X_n\}$를 n개의 범주형 데이터 집합이라 하면, 데이터 X_j $(1\le j\le n)$는 범주형 속성 $A_1, A_2, .., A_p$로 정의된다. 각 A_l은 도메인 $DOM(A_l)=\{a_l^{(1)}, a_l^{(2)}, .., a_l^{(n_l)}\}$로 표현된 값들 중 하나를 가지며, 여기서 n_l은 속성 A_l $(1\le l\le p)$이 가질 수 있는 값

의 개수이다. X_j는 $[x_{j1}, x_{j2}, \ldots, x_{jp}]$로 표현되고, 따라서 각 X_j는 속성-값 쌍의 논리적 조합 $[A_1 = x_{j1}]\&\ldots\&[A_p = x_{jp}]$로 나타낼 수 있다.

Fuzzy k-modes 알고리즘은 다음의 목적 함수 J_m을 최소화함으로써 X를 k개의 군집으로 분할하게 된다.

$$J_m(U, V:X) = \sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}^{m}\, d_c(V_i, X_j) \tag{1}$$

여기서

$$0 \le \mu_{ij} \le 1, \quad 1 \le i \le k, \quad 1 \le j \le n$$

$$\sum_{i=1}^{k}\mu_{ij} = 1, \quad 1 \le j \le n$$

$$0 < \sum_{j=1}^{n}\mu_{ij} < n, \quad 1 \le i \le k$$

μ_{ij}는 i번째 군집에 대한 데이터 X_j의 소속 정도를 나타내며, $(k \times n)$ 패턴 행렬 $U = [\mu_{ij}]$의 원소가 된다. $V = \{V_1, \ldots, V_n\}$는 퍼지 군집의 중심 벡터이다. 중심 V_i는 $[v_{i1}, v_{i2}, \ldots, v_{ip}]$로 표현된다. m은 각 데이터의 퍼지 소속 정도를 조절하는 매개 변수이다.

범주형 데이터를 분류하기 위해서, fuzzy k-modes 알고리즘은 fuzzy c-means 타입의 기법에 기반하여 k-modes 알고리즘을 확장하였다. 이를 위해 군집 중심과 데이터간의 거리 척도가 필요하며, 알고리즘 각 단계에서 군집 중심을 갱신하는 방법이 정의된다.

군집 중심 V_i와 데이터 X_j사이의 거리 척도 $d_c(V_i, X_j)$는 다음과 같이 정의된다.

$$d_c(V_i, X_j) = \sum_{l=1}^{p}\delta(v_{il}, x_{jl}) \tag{2}$$

여기서

$$\delta(v_{il}, x_{jl}) = \begin{cases} 0, & v_{il} = x_{jl} \\ 1, & v_{il} \ne x_{jl} \end{cases} \tag{3}$$

척도 $d_c(V_i, X_j)$는 범주형 데이터 집합 상에서 metric임이 알려져 있으며, 또한 일반화된 해밍 거리(Hamming distance)라고 볼 수 있다 [12].

군집 중심은 다음과 같이 갱신된다. X가 속성 $A_1, \ldots, A_p$와 도메인 $DOM(A_l) = \{a_l^{(1)}, a_l^{(2)}, \ldots, a_l^{(n_l)}\}$으로 표현된다고 가정하자. 여기서 n_l은 속성 A_l이 가질 수 있는 범주의 종류이다. 군집 중심 $V_i = [v_{i1}, \ldots, v_{ip}]$가 주어진 경우, 각 $v_{il} \in V\ (1 \le l \le p)$의 갱신은 아래와 같다.

$$v_{il} = a_l^{(r)} \in DOM(A_l) \tag{4}$$

여기서

$$\sum_{x_{jl} = a_l^{(r)}}\mu_{ij}^{m} \ge \sum_{x_{jl} = a_l^{(t)}}\mu_{ij}^{m}, \quad 1 \le t \le n_l \tag{5}$$

군집 중심 V_i의 속성 A_l은 i번째 군집에 대한 소속정도 μ_{il}값을 최대로 갖는 범주값(category value)에 의해 결정된다.

3. 제안하는 퍼지 군집화 알고리즘

3.1 접근 방향

본 논문에서는 퍼지화된 중심 표현을 갖는 퍼지 군집화 알고리즘을 제안한다. 제안하는 군집의 중심을 단일 스칼라 값으로 표현하는 대신, 퍼지 집합으로 표현함으로써 불확실한 영역에서의 모호성을 최소화한다. 기존의 퍼지 군집화 알고리즘의 경우, 현재 군집의 중심 정보가 다음 단계에서 충실히 전달되지 않는 경우가 발생한다.

예를 들어, 도메인 $DOM(A_l) = \{high, low\}$과 세 데이터 X_1, X_2, X_3가 주어진 경우 (i번째 군집에 대한 소속정도는 $\mu_{i1} = 0.70, \mu_{i2} = 0.80, \mu_{i3} = 0.15$),

$$X_1 = [x_{11}, \ldots, high, \ldots, x_{1p}]$$
$$X_2 = [x_{21}, \ldots, low, \ldots, x_{2p}]$$
$$X_3 = [x_{31}, \ldots, high, \ldots, x_{3p}]$$

군집 중심 $V = [v_{i1}, v_{i2}, \ldots, v_{ip}]$의 i-속성(v_{il})을 살펴보자. 수식 4와 5에 의해, v_{il}은 "high" 값을 가지게 된다. 이는 "high" 값이 "low" 값에 비해 더 높은 소속 정도의 합을 가지기 때문이다. 따라서 "low"는 경쟁력 있는 높은 소속 값을 가짐에도 불구하고 알고리즘의 다음 단계에서는 고려의 대상에서 제외된다. 이것이 불확실성이 있는 데이터에 적용될 경우, 군집화를 수행하는 과정에서 오류를 발생시킬 수 있다. 이러한 사항을 해결하기 위해서 본 논문에서는 퍼지화된 중심 표현을 제안하고 신뢰성있는 다양한 중심 정보를 알고리즘 각 수행 단계에서 최대한 유지한다.

기존의 중심 표현에서는 중심의 각 속성이 단일의 범주값만 가진다. 이에 반해 퍼지 중심 표현에서는 각 속성이 퍼지 범주 집합으로 정의된다. 도메인 $DOM(A_l) = \{a_l^{(1)}, a_l^{(2)}, \ldots, a_l^{(n_l)}\}$에 대해서, 제안하는 퍼지 중심 V는 다음과 같이 정의된다.

$$V = [v_1, \ldots, v_l, \ldots, v_p] \tag{6}$$

여기서

$$v_l = a_l^{(1)}/w_l^{(1)} + a_l^{(2)}/w_l^{(2)} + \ldots + a_l^{(n_l)}/w_l^{(n_l)} \tag{7}$$

$$0 \le w_l^{(t)} \le 1, \quad 1 \le t \le n_l$$

$$\sum_{t=1}^{n_l} w_l^{(t)} = 1, \quad 1 \le l \le p$$

각 속성 $v_l \in V$는 퍼지 집합 $\{(a_l^{(t)}, w_l^{(t)})\}\ (1 \le t \le n_l)$로 표현된 퍼지 범주값이며, 군집에 속한 데이터의 분포로써 결정된다. $w_l^{(t)}$는 $a_l^{(t)}$가 v_l에 대한 신뢰도를 나타낸다.

3.2 거리 척도와 중심 갱신

퍼지중심 $V = [v_1, \ldots, v_p]$와 데이터 $X = [x_1, \ldots, x_p]$간의 거리 척도는 아래와 같이 정의된다.

$$d(V, X) = \sum_{l=1}^{p}\delta(v_l, x_l) \tag{8}$$

여기서

$$\delta(v_l, x_l) = \sum_{t=1}^{n_l} r(a_l^{(t)}, x_l) \tag{9}$$

그리고

$$\tau(a_l^{(t)}, x_l) = \begin{cases} 0, & a_l^{(t)} = x_l \\ w_l^{(t)}, & a_l^{(t)} \neq x_l \end{cases} \quad (10)$$

δ-함수는 $a_l^{(t)} \in DOM(A_l)$과 x_l 사이의 거리를 합산함으로써 구할 수 있다. τ-함수는 두 값이 동일하면 0.0을, 그렇지 않을 경우 그 신뢰도 값을 반환한다. 수식 8에 의해 패턴 행렬 $U=[\mu_{ij}]$가 결정되면, 퍼지 중심은 다음과 같이 갱신된다. 각 퍼지 중심 $V_i=[v_{i1}, \ldots, v_{il}, \ldots, v_{ip}]$가 주어진 경우, 수식 7의 속성 v_{il}는 $w_l^{(t)}$ $(1 \le t \le n_l)$를 계산함으로써 갱신된다.

$$w_l^{(t)} = \sum_{j=1}^{n} \gamma(x_{jl}) \quad (11)$$

여기서

$$\gamma(x_{jl}) = \begin{cases} \mu_{ij}^m, & a_l^{(t)} = x_{jl} \\ 0, & a_l^{(t)} \neq x_{jl} \end{cases} \quad (12)$$

각 $v_{il} \in V_i$는 도메인 $DOM(A_l)$의 범주값의 분포 정보를 가지게 되며, 수식 7의 조건들이 만족함을 알 수 있다. 제안한 퍼지 중심을 앞서 살펴본 예제에 적용할 경우, 속성 $v_{il} \in V_i$ $(m=1.0)$의 값은 아래와 같다.

$$v_{il} = high/0.85 + low/0.80 \quad (13)$$

v_{il}는 각 군집에 대한 범주값들의 신뢰도를 표현하고 있으며, 알고리즘의 각 단계에서 $w_l^{(t)}$를 갱신함으로써 결정할 수 있다.

3.3 퍼지 중심을 갖는 퍼지 군집화 알고리즘

퍼지 중심을 갖는 목적 함수를 최소화하기 위해서

$$J_m(U, V:X) = \sum_{i=1}^{k} \sum_{j=1}^{n} \mu_{ij}^m \, d(V_i, X_j) \quad (14)$$

제안하는 알고리즘은 fuzzy c-means 알고리즘의 형태로 확장되었다.

Step 1. 군집의 수 k와 매개변수 m이 주어진 경우, 초기 중심값 $V(0)$ $(t=0)$을 설정한다. 각 $v_{il} \in V_i$은 $w_l^{(t)}$에 무작위 값을 할당함으로써 결정된다.

Step 2. i번째 군집을 계산한다 $(i=1,2,\ldots,k)$. 각 x_j에 대해서:

$$\mu_{ij}(t) = \left(\sum_{z=1}^{k} \left(\frac{d(V_i, X_j)}{d(V_z, X_j)} \right)^{\frac{1}{m-1}} \right)^{-1} \quad (15)$$

Step 3. 퍼지 중심 $V_i(t+1)=[v_{i1}, \ldots, v_{il}, \ldots, v_{ip}]$를 갱신한다. $v_{il}=\{(a_l^{(t)}, w_l^{(t)})\}$ $(1 \le l \le p)$에 대해서

$$w_l^{(t)} = \sum_{j=1}^{n} \gamma(x_{jl}) \quad (16)$$

여기서

$$\gamma(x_{jl}) = \begin{cases} \mu_{ij}^m, & a_l^{(t)} = x_{jl} \\ 0, & a_l^{(t)} \neq x_{jl} \end{cases} \quad (17)$$

Step 4. J_m값에 향상이 없다면 알고리즘을 중단한다; 그렇지 않으면, Step 2로 간다 ($t \leftarrow t+1$).

4. 실험 결과 및 분석

본 장에서는 제안된 알고리즘의 성능을 평가한다. 이를 위해 기존의 대표적인 방법들인 k-modes 알고리즘과 fuzzy k-modes 알고리즘과의 비교 실험을 수행하였다. k-modes 알고리즘과 fuzzy k-modes 알고리즘의 초기 중심값은 무작위로 선택되었다. 실험에 사용한 데이터 집합은 SOYBEAN [7], CREDIT [13], 그리고 ZOO [15] 집합이다. 군집화 결과는 Huang이 제안한 정확도 척도(r)로써 제시하였다 [7].

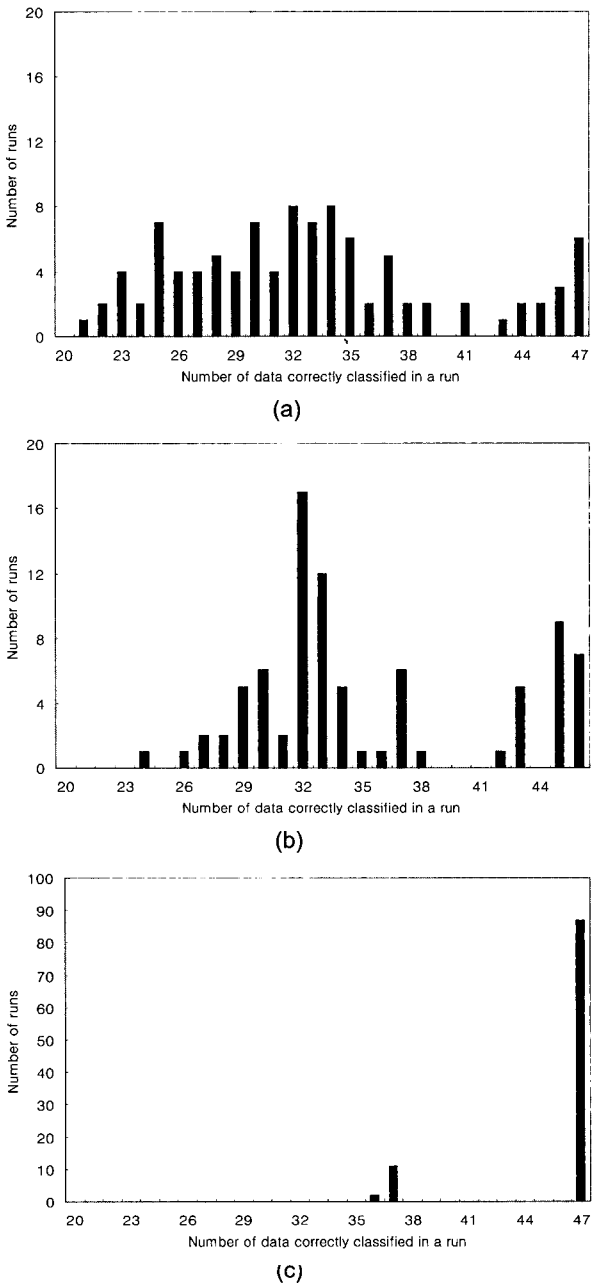

그림 1. 정확하게 분류된 데이터 수와 수행 빈도. (a) k-modes; (b) fuzzy k-modes; (c) 제안한 알고리즘
Fig. 1. Distributions of the number runs with respect to the number of correctly classified records in each run. (a) k-modes; (b) fuzzy k-modes; (c) proposed

$$r=\frac{\sum_{i=1}^{k} a_i}{n} \tag{18}$$

여기서 a_i는 i번째 군집과 실제 정답 군집에 모두 나타나는 정확하게 분류된 데이터의 수를 나타내며, n은 전체 데이터 수를 의미한다. 따라서, r값이 높을수록 좋은 군집화 결과라고 볼 수 있으며, 완벽하게 분류했을 경우 $r=1.0$의 값을 가진다.

4.1 군집화 성능 비교

SOYBEAN 데이터 집합은 질병에 대한 47개의 데이터를 가진다 [7]. 각 데이터는 35개의 범주형 속성으로 표현되며, 4가지의 군집 유형 중 하나로 분류된다: Diaporthe Stem Canker, Charcoal Rot, Rhizonctonia Root Rot, Phytophthora Rot. Phytophthora Rot은 17개의 데이터를 가지는 군집이며, 나머지 세 개의 군집은 각 10개의 데이터를 가진다. 앞에서 언급한 세 가지의 군집화 알고리즘을 대상으로 각 100번의 수행을 실시하였다 ($k=4$). 그림 1은 각 알고리즘에 대해서 정확하게 분류된 데이터 수와 수행 빈도와의 관계를 보인 것이다. k-modes와 fuzzy k-modes 알고리즘과 비교해 볼 때, 제안된 알고리즘이 우월한 성능을 보임을 알 수 있다 (100번의 수행 중 87번의 수행이 정확한 분류 결과를 제공하였다). 표 1은 각 알고리즘에 대해서 100회 수행한 결과의 평균 정확도를 제시한 것이다 ($m\in[1.1,2.0]$). Huang이 지적한 바와 같이 [7], fuzzy k-modes 알고리즘은 $m=1.1$ ($r=0.772$)에서 가장 좋은 결과를 제공함을 알 수 있었으며, k-modes 알고리즘 ($r=0.685$) 보다 우수한 결과를 보였다. 제안된 알고리즘의 경우 $m=1.8$에서 $r=0.972$의 정확도를 보였으며, fuzzy k-modes 알고리즘보다 20% 성능 향상을 보임을 알 수 있었다.

표 1. SOYBEAN 데이터 집합에 대한 클러스터링 알고리즘들의 평균 정확도 비교

Table 1. Average clustering accuracy achieved by three clustering methods for the SOYBEAN set

m	K-modes	Fuzzy k-modes	Proposed
1.1		**0.772**	0.893
1.2		0.766	0.920
1.3		0.733	0.946
1.4		0.740	0.967
1.5		0.713	0.972
1.6	**0.685**	0.693	0.955
1.7		0.694	0.964
1.8		0.703	**0.972**
1.9		0.703	0.958
2.0		0.690	0.900

CREDIT 데이터 집합은 지원자들에 대한 202개의 신용도 승인에 관한 데이터를 가진다. 각 데이터는 9개의 속성을 가지며, "승인" 또는 "거절"의 2가지 군집으로 분류된다. 각 알고리즘을 역시 100회 수행한 결과, SOYBEAN 데이터 집합과 유사한 결과를 보임을 알 수 있었다. 제안된 알고리즘은 $m=1.8$에서 $r=0.800$의 정확도를 보였다 (표 2). K-modes와 fuzzy k-modes 알고리즘은 이보다 낮은 $r=0.658$와 $r=0.744$를 각각 나타내었다. 제안된 알고리즘이 5.6%의 성능 향상을 보였다. ZOO 데이터 집합은 18개의 범주형 속성을 갖는 101개의 데이터로 구성된다. 각 데이터는 동물의 유형을 나타내는 특징들로써 전체 7가지의 군집으로 나뉜다. 100회 수행 결과, k-modes와 fuzzy k-modes 알고리즘은 각각 $r=0.602$와 $r=0.642$의 정확도를 얻을 수 있었다. 이에 반해, 제안된 알고리즘은 $m=1.8$에서 $r=0.751$의 정확도로서 fuzzy k-modes 알고리즘에 비해 10.9% 향상을 보였다.

표 2. CREDIT과 ZOO 데이터 집합에 대한 클러스터링 알고리즘들의 평균 정확도 비교

Table 2. Average clustering accuracy achieved by three clustering methods for the CREDIT and ZOO data sets

Data set	K-modes	Fuzzy k-modes	Proposed
CREDIT	0.658	0.744	**0.800**
ZOO	0.602	0.642	**0.751**

SOYBEAN, CREDIT 그리고 ZOO 데이터 집합을 이용한 실험 결과, k-modes와 fuzzy k-modes 알고리즘은 유사한 성능을 보임을 알 수 있었다. 또한 제안된 알고리즘이 위 두 알고리즘보다 우수한 성능을 제공하였다.

4.2 군집 경계 영역의 분류

군집화를 하는 데 있어서 가장 어려운 점 중 하나는 데이터들의 경계 영역, 즉 각 군집의 이웃 군집과 접하는 영역에서의 효과적인 분류를 결정하는 것이다. 이와 같은 경계 영역의 데이터들은 인접하는 이웃 군집들에 대해서 거리 차이가 크지 않기 때문에 오분류 될 가능성이 높다. 위에서 살펴본 세 가지 군집화 알고리즘의 분류 성능을 좀 더 자세히 살펴보기 위해서, 본 절에서는 SOYBEAN 데이터 집합에서 추출한 4개의 경계 데이터를 분석한다. 세 알고리즘은 같은 초기 군집 중심을 가지고 실행된다. 표 3은 세 알고리즘에 의한 4개의 데이터의 분류 결과를 나타낸 것이다. 표에는 데이터와 군집 중심 간의 거리를 제시하였으며, 각 알고리즘에 의해 판별된 군집과 실제 정답 군집을 나타내었다. 오분류된 데이터는 (*)로 표현하였다.

K-mode 알고리즘은 4가지 데이터 X_3, X_{23}, X_{25} 그리고 X_{29}를 모두 오분류하는 결과를 보였다. 이들 중 2개의 데이터는 fuzzy k-modes 알고리즘에서는 정확하게 분류되었다. 하지만, Huang이 지적한 바와 같이 [7], fuzzy k-modes 알고리즘은 우연적으로 이러한 결과를 제시한 것으로 보인다. 왜냐하면 데이터 X_3을 살펴보면, 첫 번째 군집과 세 번째 군집에 동일한 거리를 가지게 되는데, 이 경우 fuzzy k-modes 알고리즘은 무작위적로 첫 번째 군집에 할당하는 방식을 택하기 때문이다. 데이터 X_{25}도 역시 같은 관점에서 이해할 수 있다. 이에 반해, 제안한 알고리즘은 4가지 데이터 모두 정확하게 분류함을 알 수 있다. 이를 통해 주어진 데이터와 퍼지 군집 중심 사이의 거리 척도가 기존의 알고리즘에 비해 보다 세밀하고 효과적이라는 것을 알 수 있다.

표 3. 경계 영역의 데이터와 군집 중심 간의 거리, 알고리즘에 의해 할당된 군집 번호 및 정답 군집
Table 3. Distance measure between boundary data and cluster centroids, and the class no. of the assigned cluster and true cluster.

Methods	Data	Distance to Centroids				Cluster No. assigned	True class
		V1	V2	V3	V4		
K-modes	X3 (*)	6	15	**4**	12	3	1
	X23(*)	10	16	12	**7**	4	3
	X25(*)	11	16	9	**7**	4	3
	X29(*)	13	15	12	**9**	4	3
Fuzzy k-modes	X3	**6**	15	**6**	12	1	1
	X23(*)	**10**	16	11	**10**	1	3
	X25	11	16	**9**	**9**	3	3
	X29(*)	10	17	11	**7**	4	3
Proposed	X3	**6.86**	12.94	11.43	11.43	1	1
	X23	10.70	15.27	**8.32**	8.36	3	3
	X25	9.99	14.34	**7.64**	7.67	3	3
	X29	11.31	15.25	**10.13**	10.17	3	3

5. 결론

Fuzzy k-modes 알고리즘은 범주형 데이터를 효율적으로 분류한다고 알려져 있다. 그러나 군집 중심을 표현하는 단일 스칼라 값과 단순한 거리 척도를 사용함으로써 불확실성이 높은 데이터를 분류하는데는 한계점을 지닌다. 이러한 단점을 해결하기 위해, 본 논문에서는 퍼지 중심 표현을 갖는 새로운 퍼지 군집화 알고리즘을 제안하였다. 퍼지 중심은 각 속성을 퍼지 집합으로 표현하고 신뢰도를 가지도록 하였다. 그리고 이를 활용하기 위한 거리 척도와 중심 갱신 기법을 소개하였다. Fuzzy k-modes와의 비교 실험 결과 제안된 방법론이 우수한 분류 결과를 제시하였다.

참 고 문 헌

[1] J.C. Gower, "A general coefficient of similarity and some of its properties", BioMetrics, vol. 27, pp. 857-874, 1971.

[2] K.C. Gowda, E. Diday, "Symbolic clustering using a new dissimilarity measure", Pattern Recognition, vol. 24, no. 6, pp. 567-578, 1991.

[3] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data-An Introduction to Cluster Analysis. New York:Wiely Publishers, 1990.

[4] R.S. Michalski, R.E. Stepp, "Automated construction of classification: Conceptual clustering versus numerical taxonomy", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 5, pp. 396-410, 1983.

[5] M.A. Woodbury, J.A. Clive, "Clinical pure types as a fuzzy partition", J. Cybern., vol 4-3, pp. 111-121, 1974.

[6] Z. Huang, "Extensions to the k-modes algorithm for clustering large data sets with categorical values", Data Mining Knowledge Discovery, vol. 2, no. 3, 1998.

[7] Z. Huang, M.K. Ng, "A fuzzy k-modes algorithm for clustering categorical data", IEEE Transactions on Fuzzy Systems, vol. 7, no. 4, 1999.

[8] A.K. Jain, R.C. Dubes, Algorithms for Clustering, NJ:Prentice-Hall, 1998.

[9] A.K. Jain, M.N. Murty, P.J. Flynn, "Data clustering: a review", ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[10] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum: New York, 1981.

[11] J.C. Bezdek, et al, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Boston:Kluwer Academy Publishers, 1999.

[12] T. Kohonen, Content-Addressable Memories, Berlin:Springer-Verlag, 1980.

[13] J.R. Quilan, R. Quilan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1992.

[14] I.H. Witten, B.A. MacDonald, "Using concept learning for knowledge acquistion", International Journal of Man-Machine Studies, vol. 27, pp. 349-370, 1988.

[15] R. Forsyth, Zoo database in the UCI KDD Archive. [Available online] http://kdd.ics.uci.edu/, 1990.

[16] H. Lee-Kwang, K.M. Lee, "Fuzzy hypergraph and fuzzy partition", IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, vol. 25, no. 2, pp. 196-201, 1995.

저 자 소 개

김대원(Dae-Won Kim)
제 10권 1호 참조

Phone : 042-869-4356
Fax : 042-869-8680
E-mail : dwkim@if.kaist.ac.kr

이광형(Kwang H. Lee)
제 10권 1호 참조

Phone : 042-869-4313
Fax : 042-869-8680
E-mail : khlee@if.kaist.ac.kr