중첩성과 분리성을 이용한 퍼지 클러스터 평가척도

A Fuzzy Cluster Validity based on Inter-cluster Overlapping and Separation

김대원*, 이광형
한국과학기술원 전자전산학과

Dae-Won Kim* and Kwang H. Lee
Department of Electrical Engineering and Computer Science,KAIST
E-mail : dwkim@if.kaist.ac.kr

ABSTRACT

본 논문에서는 퍼지 클러스터링 알고리즘에 의해 구해진 퍼지 클러스터들에 대한 평가척도를 제안한다. 제안된 척도는 퍼지 클러스터들간의 중첩성(overlapping)과 분리성(separation)을 이용한다. 중첩성은 클러스터간 인접도를 이용하여 계산하며, 분리성은 데이터에 대한 상관성 정도를 나타낸다. 따라서 중첩성이 낮고 분리성이 높을수록 좋은 클러스터 결과라고 할 수 있다. 표준 데이터 집합을 대상으로 기존의 척도들과 비교실험함으로써 제안된 척도의 신뢰성을 알아보았다.

I. 서론

퍼지 클러스터링 알고리즘의 목적은 주어진 데이터 집합을 주어진 수의 유사한 퍼지 클러스터로 분할하는 것이다. 지금까지 가장 널리 사용되는 퍼지 클러스터링 알고리즘은 Bezdek에 의해 제안된 Fuzzy C-Means (FCM) 알고리즘이다 [1]. FCM 알고리즘의 목적은 주어진 데이터 집합 $X = \{x_1, ..., x_n\}$와 분할하고자 하는 클러스터의 수 c에 대해서 아래의 함수 J_m을 최소화함으로써 퍼지 분할 $\tilde{F} = \{\tilde{F}_1, \tilde{F}_2, .., \tilde{F}_c\}$를 구하는 것이다.

$$J_m(U, V : X) = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^m \|x_j - v_i\|^2 \qquad (1)$$

여기서 μ_{ij} 는 퍼지 클러스터 $\tilde{F}_i$에 대한 데이터 x_j의 소속정도를 나타내며, $(c \times n)$ 패턴 행렬 $U = [\mu_{ij}]$의 원소가 된다. $V = (v_1, v_2, .., v_c)$ 는 퍼지 클러스터들의 중심 벡터의 집합이다. $\|x_j - v_i\|^2$ 는 데이터 x_j와 클러스터 중심 v_i간의 유클리디언 거리를 나타낸다. 매개변수 m은 각 데이터의 소속정도에 대한 퍼지값을 조종하는 역할을 한다. 일반적으로 $m = 2.0$의 설정이 좋은 결과를 제공한다고 알려져 있다 [4].

그러나 이 알고리즘은 클러스터의 중심(centroid)을 초기화하는 문제 때문에 최적의 분할 결과를 얻는데 어려움이 있다. 대부분의 클러스터링 알고리즘은 초기화를 무작위 값으로 선정하기 때문에, 초기 클러스터 중심값의 변화는 이후 얻어지는 클러스터 분할 결과에 많은 영향을 끼치게 된다. 따라서, 퍼지 클러스터 분할이 판별되고 나면, 이를 평가할 수 있는 방법이 필요하게 되었다. 이것을 가능케 하는 것이 평가척도(validity index)이다. 더욱이, 평가척도를 사용함으로써 클러스터 분할의 개수를 미리 알지 못하는 상황에서, 최적의 클러스터 수를 찾을 수 있다 [8]. 지금까지 다양한 퍼지 클러스터 평가척도가 제안되어져 왔다.

Bezkek은 퍼지 클러스터링을 위해서 두가지의 클러스터 평가 척도를 제안하였다 [2][3]: 분할 계수(v_{PC})와 분할 엔트로피 (v_{PE}). v_{PC}는 최대값을 가질때 최적 분할을 산출하며, v_{PE}는 최소가 되는 지점에서 최적의 분할 결과를 제공한다. Xie와 Beni는 두가지 개념(조밀성과 분리성)에 초점을 맞춘 평가척도를 제안하였다 (v_{XB}) [5]. Fukuyama와 Sugeno는 또한 클러스터 내부의 조밀성과 클러스터 중심과의 거리를 이용한 분할 결과를 평가하였다 (v_{FS}) [6]. Kwon은 Xie-Beni의 방법을 확장하여 기존 척도의 단조감소 경향을 회피하려고 시도하였다 (v_K) [7].

최근에는 클러스터 내부의 분산정도를 고려한 척도가 제안되고 있다. Rezaee는 클러스터 내부의 분산과 거리 함수를 병합한 척도를 제안하였다 (v_{CWB}) [8], Boudraa도 역시 유사한 접근법에 기반한 평가 척도를 제안하였다 ($v_{B_{crit}}$) [9]. Zahid는 퍼지 조밀성과 퍼지 분리성을 제안하고 이를 기존의 방법들과 병합하여 사용하였다 (v_{SC}) [10]. 퍼지 조밀성과 퍼지 분리성을 계산하기 위해서 퍼지 합집합과 교집합 연산을 각각 응용하였다. Kim은 최적 클러스터의 수를 결정하기 위하여 고밀도 분할척도와 저밀도 분할 척도라는 개념을 도입하기도 하였다 (v_{SV}) [11].

II. 제안한 퍼지 클러스터 평가척도

2.1 연구배경

서론에서 살펴본 기존의 평가척도들은 기하학적 해석을 하는 데 있어 한계점을 지닌다. 이것은 대부분의 척도들이 클러스터 중심간의 거리만을 이용해서 조밀성(compactness)과 분리성(separation)을 계산해왔기 때문이다.

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center

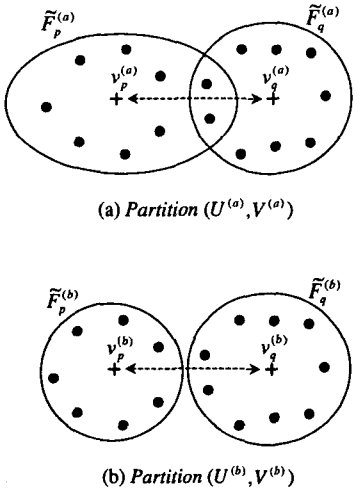

Fig. 1. 동일한 분리도를 갖는 두 퍼지 분할 $(U^{(a)},V^{(a)})$와 $(U^{(b)},V^{(b)})$

먼저 조밀성에 대해 살펴보면, 대부분의 기존 척도들은 $\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\|x_j-v_i\|^2$와 같은 방식을 이용해서 조밀성을 계산한다 [5][6][8][9]. 그러나 Pal [4]과 Kwon [7]이 지적하였듯이, 이런 방식은 클러스터의 수(c)가 커질수록 평가 값이 단조 감소하는 경향을 보이게 된다. 따라서 클러스터의 수가 매우 큰 경우에는 적용하기가 어렵다. 클러스터 중심을 이용한 분리성 계산도 역시 한계점을 지닌다 [5][7][8][9]. 그림 1은 이와 같은 기존 방법들의 문제점을 잘 보여준다. 즉, 클러스터 중심 사이의 거리가 같은 두개의 퍼지 분할 $(U^{(a)},V^{(a)})$와 $(U^{(b)},V^{(b)})$을 나타낸 것이다. 그림 1(a)에는, 두개의 퍼지 클러스터 $\tilde{F}_p^{(a)},\tilde{F}_q^{(a)}\in U^{(a)}$와 그들의 중심 $v_p^{(a)}$, $v_q^{(a)}\in V^{(a)}$가, 그리고 그림 1(b)에는 중심 $v_p^{(b)},v_q^{(b)}\in V^{(b)}$를 가지는 또 다른 두 개의 퍼지 클러스터가 존재한다. 직관적으로 $(U^{(b)},V^{(b)})$가 $(U^{(a)},V^{(a)})$보다 잘 분할되었다는 것을 알 수 있다. 하지만, 기존의 척도로는 중심사이의 거리 $\|v_p^{(a)}-v_q^{(a)}\|$와 $\|v_p^{(b)}-v_q^{(b)}\|$가 동일하기 때문에 두 경우의 분할을 구분할 수가 없다.

본 논문에서는 이러한 기존 척도들의 문제점을 극복하기 위해서 클러스터 중심간의 거리를 이용하는 대신 전체 클러스터간의 중첩성(overlapping)과 분리성(separation)을 사용하게 된다. 이를 위해 각 퍼지 클러스터는 개개의 퍼지 집합으로 간주한다.

$$\tilde{F}_i=\sum_{j=1}^{n}\mu_{\tilde{F}_i}(x_j)/x_j=\mu_{\tilde{F}_i}(x_1)/x_1+\ldots+\mu_{\tilde{F}_i}(x_n)/x_n \quad (2)$$

2.2 클러스터간의 중첩성 계산

전체 클러스터간 중첩성(overlapping)을 구하기 앞서, 주어진 소속정도(μ)에 대한 클러스터간 중첩 함수를 먼저 계산한다. 두 퍼지 클러스터 $\tilde{F}_p$, $\tilde{F}_q$와 주어진 소속 정도 μ에 대해서 중첩함수(overlapping function) $f(\mu)$는 다음과 같이 정의된다:

$$f(\mu:\tilde{F}_p,\tilde{F}_q)=\sum_{j=1}^{n}\delta(x_j,\mu:\tilde{F}_p,\tilde{F}_q)\omega(x_j) \quad (3)$$

$$\delta(x_j,\mu:\tilde{F}_p,\tilde{F}_q)=\begin{cases}1.0 & \text{if } \mu\le MIN(\mu_{\tilde{F}_p}(x_j),\mu_{\tilde{F}_q}(x_j))\\ 0.0 & \text{otherwise}\end{cases} \quad (4)$$

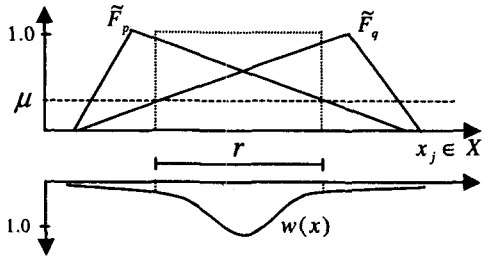

Fig. 2. 소속정도 μ에서 두 클러스터간 중첩성 $f(\mu)$

$\delta(x_j,\mu:\tilde{F}_p,\tilde{F}_q)$는 데이터 x_j에 대해 소속정도 μ에서 두 클러스터가 중첩하는지를 결정한다. 두 클러스터 모두 μ 값 이상이면 중첩성은 1.0의 값을 가진다. 그렇지 않은 경우에는 0.0의 값을 가진다. 또한, 모호한 데이터에 대한 가중치를 할당하기 위해서 $\omega(x_j)$ 함수를 도입하였다. 가중치 $\omega(x_j)\in[0.0,1.0]$ 는 두 클러스터간의 공유 정도에 따라 상대적으로 적용된다. 이러한 가중치 적용은 클러스터간에 많이 중첩된 모호한 데이터 판별에 있어서 장점을 가진다.

그림 2는 소속정도 μ에서 두 클러스터간의 중첩성 계산을 도식화한 것이다. 전체 데이터 $x_j\in X$에 대해서 $x_j\in r$만이 식 4에 의해서 1.0의 중첩성을 부여받는다. 주어진 가중치 함수 $w(x)$를 이용한 인접함수 $f(\mu:\tilde{F}_p^{(a)},\tilde{F}_q^{(a)})$는 $\delta(x_j,\mu:\tilde{F}_p^{(a)},\tilde{F}_q^{(a)})$와 $w(x_j)$의 곱을 누적함으로써 계산된다.

정의 1: 패턴 행렬 U에 속하는 두 퍼지 클러스터 $\tilde{F}_p$와 $\tilde{F}_q$에 대해서, 각 소속 정도 μ에서의 중첩함수 $f(\mu:\tilde{F}_p,\tilde{F}_q)$가 주어진 경우, 두 클러스터간의 중첩성 $P(\tilde{F}_p,\tilde{F}_q)$는 다음과 같이 정의된다

$$P(\tilde{F}_p,\tilde{F}_q)=\sum_{\mu}f(\mu:\tilde{F}_p,\tilde{F}_q)=\sum_{\mu}\sum_{j=1}^{n}\delta(x_j,\mu:\tilde{F}_p,\tilde{F}_q)\omega(x_j) \quad (5)$$

$P(\tilde{F}_p,\tilde{F}_q)$는 $f(\mu:\tilde{F}_p,\tilde{F}_q)$를 전체 소속정도의 범위에 대해서 계산한 것이다. 따라서, $P(\tilde{F}_p,\tilde{F}_q)$가 상대적으로 낮은 값을 가진다는 것은 클러스터 $\tilde{F}_p$와 클러스터 $\tilde{F}_q$의 중첩성이 낮다는 것을 의미하므로, 두 클러스터가 잘 분할 되었다는 것을 알 수 있다. 두 클러스터간 중첩성 정의를 기반으로, 전체 퍼지 클러스터 분할에 대한 중첩성은 아래와 같이 정의할 수 있다.

정의 2: 두 퍼지 클러스터간의 중첩성 $P(\tilde{F}_p,\tilde{F}_q)$와 분할된 클러스터의 수 c가 주어진 경우, 전체 퍼지 클러스터의 중첩성 $Overlap(c,U)$은 다음과 같이 정의된다

$$Overlap(c,U)=\frac{1}{{}_cC_2}\underbrace{\sum_{p=1}^{c}\sum_{q=1}^{c}}_{p\neq q}P(\tilde{F}_p,\tilde{F}_q) \quad (6)$$

여기서 ${}_cC_2$는 클러스터간 중첩성 계산 수를 나타내므로, $Overlap(c,U)$는 분할에 속하는 모든 클러스터들 간의 평균 중첩성을 나타낸다. 그러므로 $Overlap(c,U)$의 값이 낮을수록 좋은 분할 결과라고 할 수 있다.

2.3 클러스터간의 분리성 계산

클러스터 중심간의 거리를 이용한 기존 척도들의 단점

을 보완하기 위해서, 본 논문에서 퍼지집합의 거리 연산을 이용한 분리성을 제안한다. 퍼지집합간의 거리를 계산하기 위해서, Lee et al이 기 제안한 유사성 척도를 활용한다 [14]. 두 퍼지집합 $\tilde{F}_p$와 $\tilde{F}_q$간의 유사성 척도 $S(\tilde{F}_p, \tilde{F}_q)$은 다음과 같이 정의된다.

$$S(\tilde{F}_p, \tilde{F}_q) = \underbrace{max}_{x \in X} min(\mu_{\tilde{F}_p}(x), \mu_{\tilde{F}_q}(x)) \quad (7)$$

위 유사성 척도는 Minkowski 거리로 대표되는 기하학적 거리에 제한받지 않으면서 두 그룹간의 상호연계 정도를 신뢰성있게 표현하고 있다 [14]. 그리고 아래 다섯 가지 속성들을 만족한다.

- $S(\tilde{F}_p, \tilde{F}_q)$는 $\tilde{F}_p \cap \tilde{F}_q$의 최대 만족도 값이다.
- $0 \leq S(\tilde{F}_p, \tilde{F}_q) \leq 1$.
- $S(\tilde{F}_p, \tilde{F}_q) = S(\tilde{F}_q, \tilde{F}_p)$.
- If $\tilde{F}_p = \tilde{F}_q$, $S(\tilde{F}_p, \tilde{F}_q) = 1$. If $\tilde{F}_p \cap \tilde{F}_q = \emptyset$, $S(\tilde{F}_p, \tilde{F}_q) = 0$.
- $\tilde{F}_p$, $\tilde{F}_q$가 일반집합일 경우, $S = 0$ if $\tilde{F}_p \cap \tilde{F}_q = \emptyset$, $S = 1$ if $\tilde{F}_p \cap \tilde{F}_q \neq \emptyset$.

예를 들어, 두 퍼지 클러스터 $\tilde{F}_p$, $\tilde{F}_q$에 대해 $S(\tilde{F}_p, \tilde{F}_q) = 0.4$의 값은 클러스터 $\tilde{F}_p$와 $\tilde{F}_q$가 적어도 0.4만큼 유사하다 또는 상호 관련이 있다는 것으로 해석할 수 있다. 다시 말해, 위 두 클러스터는 0.6의 값으로 분리되어 있다 또는 상호 관련이 없다라고 할 수 있다. 이와 같은 사실에 기반하여 제안한 전체 클러스터 분할의 분리성을 다음과 같이 정의한다.

정의 3: 패턴 행렬 U에 속하는 두 퍼지 클러스터 $\tilde{F}_p$와 $\tilde{F}_q$에 대한 유사성 척도를 $S(\tilde{F}_p, \tilde{F}_q)$ 하자. 분할된 클러스터의 수 c가 주어진 경우, 전체 퍼지 클러스터의 분리성 Sep(c,U)는 다음과 같이 정의된다.

$$Sep(c,U) = 1 - \underbrace{min}_{p \neq q} S(\tilde{F}_p, \tilde{F}_q) \quad (8)$$

$$= 1 - \underbrace{min}_{p \neq q}[\underbrace{max}_{x \in X} min(\mu_{\tilde{F}_p}(x), \mu_{\tilde{F}_q}(x))] \quad (9)$$

분리성 $Sep(c,U)$ 모든 퍼지 클러스터들 간의 최소 거리를 나타내게 된다. 그러므로 $Sep(c,U)$의 값이 클수록 좋은 분할 결과라고 할 수 있다.

2.4 중첩성과 분리성을 이용한 평가척도

앞에서 정의한 두 척도 중첩성 $Overlap(c,U)$과 분리성 $Sep(c,U)$은 서로 다른 단위를 가지기 때문에 아래와 같은 정규화 과정이 필요하다. 모든 가능한 $c = 2, 3, ..., c_{max}$에 대해서 다음과 같은 값을 가질 경우:

$$Overlap = \{Overlap(2,U), ..., Overlap(c_{max},U)\} \quad (10)$$

$$Sep = \{Sep(2,U), ..., Sep(c_{max},U)\} \quad (11)$$

각 척도에 대해서 최대값을 계산하고,

$$Overlap_{max} = \underbrace{max}_{c} Overlap(c,U) \quad (12)$$

$$Sep_{max} = \underbrace{max}_{c} Sep(c,U) \quad (13)$$

그 최대값을 이용하여 각 척도를 정규화시킨다. 정규화된 중첩성을 $Overlap^N(c,U)$로, 정규화된 분리성을 $Sep^N(c,U)$로 표기하면, 본 논문에서 제안하는 최종적인 클러스터 척도는 다음 정의 4와 같다.

$$Overlap^N(c,U) = \frac{Overlap(c,U)}{Overlap_{max}} \quad (14)$$

$$Sep^N(c,U) = \frac{Sep(c,U)}{Sep_{max}} \quad (15)$$

정의 4: 패턴 행렬 U와 클러스터의 수 c에 대해서, 전체 퍼지 클러스터 분할에 대한 정규화된 중첩성과 분리성을 각각 $Overlap^N(c,U)$와 $Sep^N(c,U)$라 하자. 그러면 제안하는 퍼지 클러스터 평가 척도 $v_{OS}(c,U)$는 다음과 같이 정의된다.

$$v_{OS}(c,U) = \frac{Overlap^N(c,U)}{Sep^N(c,U)} \quad (16)$$

$v_{OS}(c,U)$는 중첩성과 분리성의 비율로 정의된다. 적은 값의 $v_{OS}(c,U)$은 클러스터들 간의 낮은 중첩성과 높은 분리성을 나타낸다고 볼 수 있다. 따라서, 최적의 퍼지 클러스터 분할 또는 최적의 클러스터 수는 $c = 2, 3, ..., c_{max}$에 대해서 $v_{OS}(c,U)$를 최소화함으로써 구할 수 있다.

III. 성능 비교 실험

제안된 척도의 신뢰성을 보이기 위해서, 다섯가지의 표준 데이터 집합에 대해서 기존의 척도들과 성능 비교 실험을 수행하였다: Bezdek's v_{PC} [2] and v_{PE} [3], Xie and Beni's v_{XB} [5], Fukuyama and Sugeno's v_{FS} [6], Kwon's v_K [7], Razaee's v_{CWB} [8], Boudraa's $v_{B_{crit}}$ [9], Zahid's v_{SC} [10], and Kim's v_{SV} [11]. 각 실험 집합에 대해서, 표준 FCM 알고리즘이 산출한 분할 결과를 $c = 2, .., c_{max}$ 범위에서 평가하였다. 사용된 데이터 집합은 BENSAID 데이터($c = 3$) [12], STARFIELD 데이터($c = 8$) [5], IRIS 데이터($c = 2$) [4], X30 데이터($c = 3$) [13], BUTTERFLY데이터($c = 2$) [7]이다. 본 논문에서는 지면 관계상 대표적인 두개의 데이터 집합의 결과를 기술하며, 나머지 데이터 집합에 대한 결과는 아래의 요약 테이블로 제시하였다.

테이블 I은 STARFIELD 데이터 집합에 대한 각 인덱스들의 평가값을 기술한 것이다. STARFIELD 집합은 66개의 데이터를 가지며, 최적의 분할 클러스터의 수는 8 또는 9로 알려져 있다. v_{CWB}와 v_{OS}가 정확히 8-클러스터로 분할한 경우가 최적이라고 계산하였다. 이와 달리, v_{PC}, v_{PE}, v_{SV}는 2개의 클러스터가 최적이라는 결과를, v_{XB}와 v_K는 최적 클러스터가 6으로, $V_{B_{crit}}$는 최적 분할로 5를 제시하였다. 그리고 v_{FS}는 $c = 7$을, v_{SC}는 $c = 3$을 최적으로 계산하였다. 테이블 II는 BENSAID 데이터 집합에 대한 평가값을 나타낸 것으로, 보는 바와 같이 v_{PC}, v_{XB}, v_K, $v_{B_{crit}}$, v_{OS}가 정확한 평가 결과를 제시한다. 나머지 척도들은 최적의 c를 찾지 못하였다.

테이블 III는 다섯 데이터 집합에 대해서 각 평가 척도들이 계산한 최적 클러스터 수를 표시한 것이다. 결과에서 보는 바와 같이 제안된 인덱스 $V_{proposed}$는 모든 데이터 집합에 대해서 정확한 평가 결과를 제시한다. 이에 반해, 대부분의 기존 척도들은 앞서 살펴본 STARTFIELD 결과에서와 같이 데이터 집합에 따라 불안정한 결과를 제시함을 알 수 있다.

TABLE I

STARFIELD 데이터 집합에 대한 평가 결과 ($c = 2, .., c_{max} = \sqrt{n} \approx 8$)

c	v_{PC}	v_{PE}	v_{XB}	v_{FS}	v_K	v_{CWB}	$v_{B_{crit}}$	v_{SC}	v_{SV}	v_{OS}
c = 2	**0.73**	**0.18**	0.24	216235.05	16.04	0.17	4.90	-0.14	**0.00**	1.25
c = 3	0.66	0.26	0.12	-597582.27	8.29	0.12	4.23	**-0.04**	0.55	0.69
c = 4	0.62	0.32	0.12	-835444.94	8.74	0.10	4.19	-0.67	0.90	0.71
c = 5	0.63	0.33	0.11	-1047072.42	8.16	0.09	**4.09**	-0.64	1.15	0.60
c = 6	0.65	0.33	**0.10**	-1266918.83	**8.09**	0.08	4.30	-0.54	1.38	0.46
c = 7	0.66	0.33	0.11	**-1394217.46**	9.61	0.07	4.66	-0.45	1.72	0.43
c = 8	0.67	0.33	0.12	-1368962.28	10.42	**0.07**	5.10	-0.55	1.81	**0.34**

TABLE II

BENSAID 데이터 집합에 대한 평가 결과 ($c = 2, .., c_{max} = \sqrt{n} \approx 7$)

c	v_{PC}	v_{PE}	v_{XB}	v_{FS}	v_K	v_{CWB}	$v_{B_{crit}}$	v_{SC}	v_{SV}	v_{OS}
c = 2	0.72	**0.19**	0.24	3671.01	11.89	0.84	8.00	-0.17	**0.00**	1.27
c = 3	**0.75**	0.20	**0.07**	-15676.31	**4.12**	0.62	**4.46**	2.56	1.01	**0.35**
c = 4	0.61	0.32	0.27	-15035.68	15.87	0.58	8.19	**4.08**	0.76	0.65
c = 5	0.66	0.29	0.12	-27285.22	8.71	0.46	7.69	4.06	1.18	0.56
c = 6	0.63	0.33	0.10	-28692.18	8.14	**0.43**	8.15	3.68	1.28	0.51
c = 7	0.61	0.36	0.10	**-29292.01**	9.20	0.44	9.09	3.35	1.49	0.55

TABLE III

다섯 데이터 집합에 대한 각 척도들의 최적 클러스터 수

Data set	$c_{optimal}$	v_{PC}	v_{PE}	v_{XB}	v_{FS}	v_K	v_{CWB}	$v_{B_{crit}}$	v_{SC}	v_{SV}	v_{OS}
X30	3	**3**	**3**	**3**	4	2	**3**	**3**	5	**3**	**3**
BENSAID	3	**3**	2	**3**	7	**3**	6	**3**	4	2	**3**
STARFIELD	8	2	2	6	7	6	**8**	5	3	2	**8**
IRIS	2	**2**	**2**	**2**	3	**2**	3	3	3	4	**2**
BUTTERFLY	2	**2**	**2**	**2**	3	**2**	**2**	**2**	**2**	**2**	**2**

IV. 결론

본 논문에서는 새로운 퍼지 클러스터 평가 척도를 제안하였다. 제안된 척도는 기존 척도들의 한계점을 극복하기 위해서 클러스터간의 중첩성과 분리성을 이용하였다. 클러스터간 중첩성이 낮을 수록, 분리성이 높을 수록 좋은 분할 결과로 판단된다. 따라서 최적의 분할 결과는 제안된 척도를 최소화시키는 방향으로 수렴된다. 비교 실험에서 살펴본 바와 같이 다양한 데이터 집합에서 기존 척도들보다 신뢰성이 높은 것으로 나타났다.

REFERENCES

[1] J.C. Bezdek (1981) Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, NY.

[2] J.C Bezdek (1974) "Numerical taxonomywith fuzzy sets", J. Math. Biology, 1:57-71.

[3] J.C Bezdek (1974) "Cluster validity with fuzzy stes", J. Cybernit., 3:58-72.

[4] N.R. Pal, J.C. Bezdek (1995) "On cluster validity for the fuzzy c-means model", IEEE Transactions on Fuzzy Systems, 3(3):370-379.

[5] X.L. Xie, G. Beni (1991) "A validity measure for fuzzy clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(8):841-847.

[6] Y. Fukuyama, M. Sugeno (1989) "A new method of choosing the number of clusters for the fuzzy c-means method", Proceedings of 5th Fuzzy Systems Symposium, 247-250.

[7] S.H. Kwon (1998) "Cluster validity index for fuzzy clustering", Electronics Letters, 34(22):2176-2177.

[8] M.R. Rezaee, B.P.F. Lelieveldt, J.H.C Reiber (1998) "A new cluster validity index for the fuzzy c-mean", Pattern Recognition Letters, 19:237-246.

[9] A.O. Boudraa (1999) "Dynamic estimation of number of clusters in data sets", Electronics Letters, 35(19):1606-1607.

[10] N. Zahid, M. Limouri, A. Essaid, "A new cluster-validity for fuzzy clustering", Pattern Recognition, vol. 32, pp. 1089-1097, 1999.

[11] D.J. Kim, Y.W. Park, D.J. Park, "A novel validity index for determination of the optimal number of clusters", IEICE Transactions on Information and Systems, E84-D(2), pp. 281-285, 2001.

[12] A.M. Bensaid, et al (1996) "Validity-guided (re)clustering with applications to image segmentation", IEEE Transactions on Fuzzy Systems, 4(2):112-123.

[13] J.C. Bezdek, N.R. Pal (1998) "Some new indexes of cluster validity", IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 28(3):301-315.

[14] H. Lee-Kwang, Y.S. Song, K.M. Lee (1994) "Similarity measure between fuzzy sets and between elements", Fuzzy Sets and Systems, 62: 291-293.