# 최적 클러스터 분할을 위한 FCM 평가 인덱스

김대원*, 이광형
한국과학기술원 전자전산학과
dwkim@if.kaist.ac.kr

# A fuzzy cluster validity index for the evaluation of Fuzzy C-Means algorithm

Dae-Won Kim* and Kwang H. Lee
Department of Electrical Engineering and Computer Science, KAIST

요 약

본 논문에서는 Fuzzy C-Means (FCM) 알고리즘에 의해 계산된 퍼지 클러스터들에 대한 평가 인덱스를 제안한다. 제안된 인덱스는 퍼지 클러스터들간의 인접성(inter-cluster proximity)을 이용한다. 클러스터 인접성을 도입함으로써 클러스터간의 중첩 정도를 계산할 수 있다. 따라서, 인접성 값이 낮을수록 클러스터들은 공간에 잘 분포하게 됨을 알 수 있다. 다양한 데이터 집합에 대한 실험을 통해서 제안된 인덱스의 효율성과 신뢰성을 검증하였다.

## 1. 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]. 하지만, FCM은 사용자가 미리 분할하고자 하는 클러스터의 수를 지정해야 하며, 주어진 클러스터 수에 따라서 매우 상이한 분할 결과를 산출한다. 그러므로 클러스터가 분할된 이후 각 퍼지 분할에 대해서 평가를 내리는 것이 필요하다. 지금까지 다양한 퍼지 클러스터 평가 인덱스가 제안되어져 왔다 [1][2][3][4][5][6][7][8][9]. Bezdek의 분할계수(partition coefficient) [2]와 분할 엔트로피(partition entropy) [3], 그리고 Xie-Beni의 인덱스 [5]가 그 대표적인 것들이다.

본 논문에서는 이러한 퍼지 클러스터링을 위한 새로운 평가 인덱스를 제안한다. 제안하는 인덱스는 클러스터간의 인접성에 기반하여 퍼지 분할 결과를 평가한다. 인접성 값은 클러스터 사이의 중첩 정도를 나타낸다. 그러므로 낮은 값의 인접성은 클러스터간의 중첩정도가 높지 않은 잘 분산된 퍼지 분할 상태를 나타낸다고 볼 수 있다.

## 2. 관련연구

FCM 알고리즘은 퍼지 분할을 얻는데 있어 다양한 분야에서 널리 이용되어 왔다. 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 평가 인덱스(validity index)이다. 더욱이, 평가 인덱스를 사용함으로써 클러스터 분할의 개수를 미리 알지 못하는 상황에서, 최적의 클러스터 수를 찾을 수 있다 [8].

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

FCM 알고리즘의 목적은 주어진 데이터 집합 $X = \{x_1, .., x_n\}$와 분할하고자 하는 클러스터의 수 $c$에 대해서 아래의 함수 $J_m$을 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 $\mu_{ij}$ 는 퍼지 클러스터 $\tilde{F}_i$에 대한 데이터 $x_j$의 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 $x_j$와 클러스터 중심 $v_i$간의 유클리디언 거리를 나타낸다. 매개변수 $m$은 각 데이터의 소속정도에 대한 퍼지값을 조종하는 역할을 한다. 일반적으로 $m = 2.0$의 설정이 좋은 결과를 제공한다고 알려져 있다 [4].

Bezkek은 퍼지 클러스터링을 위해서 두가지의 클러스터 평가 인덱스를 제안하였다 [2][3]: 분할 계수($V_{PC}$)와 분할 엔트로피 ($V_{PE}$). 식 2에서 $V_{PC}$는 최대값을 가질때 최적 분할을 산출하며, $V_{PE}$는 최소가 되는 지점에서 최적의 분할 결과를 제공한다.

$$V_{PC} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^2}{n},\ V_{PE} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}[\mu_{ij}\ log_a(\mu_{ij})] \qquad (2)$$

Xie와 Beni는 두가지 개념(조밀성과 분리성)에 초점을 맞춘 평가 인덱스를 제안하였다 ($V_{XB}$) [5]. 식 3의 분자항은 퍼지 분할의 조밀성(compactness)을 나타내며, 분모항은 클러스터간의 분리정도(separation)를 나타낸다. Fukuyama와 Sugeno는 또한 클러스터 내부의 조밀성과 클러스터 중심과의 거리를 이용한 분할 결과를 평가하였다 ($V_{FS}$) [6]. Kwon은 Xie-Beni의 인덱스를 확장하여 기존 인덱스의 단조감소 경향을 회피하려

고 시도하였다 ($V_K$) [7].

$$V_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\|x_j - v_i\|^2}{n(\underbrace{min}_{i\neq k}\|v_i - v_k\|^2)} \quad (3)$$

$$V_{FS} = \sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^2(\|x_k - v_i\|^2 - \|v_i - \bar{v}\|^2) \quad (4)$$

$$V_K = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^2\|x_j - v_i\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|v_i - \bar{v}\|^2}{\underbrace{min}_{i\neq k}\|v_i - v_k\|^2} \quad (5)$$

Rezacc는 클러스터 내부의 분산과 거리 함수를 병합한 인덱스를 제안하였다 ($V_{CWB}$) [8]. 식 6에서 $\sigma(v_i)$는 $i$-번째 퍼지 클러스터의 데이터 분산을 나타내며, $\sigma(X)$는 전체 패턴 집합 $X$의 분산을 나타낸다. Boudraa도 역시 유사한 접근법에 기반한 평가 인덱스를 제안하였다 ($V_{B_{crit}}$) [9].

$$V_{CWB} = \alpha\frac{\sum_{i=1}^{c}\|\sigma(v_i)\|}{c\|\sigma(X)\|} + \frac{D_{max}}{D_{min}}\sum_{k=1}^{c}(\sum_{z=1}^{c}\|v_k - v_z\|)^{-1} \quad (6)$$

$$V_{B_{crit}} = \frac{max\ \delta(v_i, v_j)}{\underbrace{min}_{i\neq j}\ \delta(v_i, v_j)} + \alpha\frac{1}{c}\frac{\sum_{q=1}^{P}\sum_{k=1}^{c}var_q(k)}{\sum_{q=1}^{P}var_t(q)} \quad (7)$$

## 3. 제안된 클러스터 평가 인덱스

### 3.1 접근방법

본 논문에서는 FCM 클러스터링 알고리즘의 결과로 얻어진 퍼지 분할에 대한 새로운 평가 인덱스를 제안한다. 기본 아이디어는 퍼지 클러스터의 기하학적 특성을 활용하는 것이다. 앞에서 살펴본 기존의 평가 인덱스들은 기하학적 해석을 하는 데 있어 한계점을 지닌다. 이것은 대부분의 인덱스들이 클러스터 중심간의 거리만을 그 해석의 주된 방법으로 고려해 왔기 때문이다.

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 $(U^{(a)}, V^{(a)})$가 $(U^{(b)}, V^{(b)})$보다 잘 분할되었다는 것을 알 수 있다. 하지만, 기존의 인덱스로는 중심사이의 거리 $\|v_p^{(a)} - v_q^{(a)}\|$와 $\|v_p^{(b)} - v_q^{(b)}\|$가 동일하기 때문에 두 경우의 분할을 구분할 수가 없다.

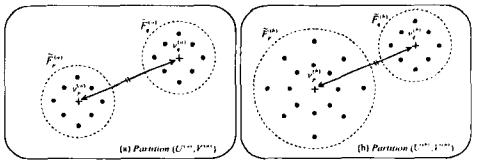

Fig. 1. 동일한 분리도를 갖는 두 퍼지 분할 $(U^{(a)}, V^{(a)})$와 $(U^{(b)}, V^{(b)})$

본 논문에서는 기존 인덱스의 문제점을 극복하기 위해서 클러스터 중심간의 거리를 이용하는 대신 전체 클러스터간의 인접성(inter-cluster proximity)을 제안한다. 이를 위해 각 퍼지 클러스터는 개개의 퍼지집합으로 표현된다.

$$\tilde{F}_i = \sum_{j=1}^{n}\mu_{\tilde{F}_i}(x_j)/x_j = \mu_{\tilde{F}_i}(x_1)/x_1 + ... + \mu_{\tilde{F}_i}(x_n)/x_n \quad (8)$$

주어진 퍼지 분할 $(U, V)$에 대해서 제안한 인덱스는 퍼지 집합 사이의 유사도를 계산함으로써 퍼지 클러스터간의 인접성을 유도한다. 이러한 인접성은 클러스터간의 중첩정도(overlap)를 의미하는 것으로 생각할 수 있으며, 낮은 값의 인접성은 클러스터간 중첩정도가 역시 낮음을 의미하기 때문에 보다 나은 분할 결과로 판단할 수 있다.

### 3.2 클러스터간의 인접성 계산

전체 클러스터간 인접성을 구하기 앞서, 주어진 소속정도($\mu$)에 대한 클러스터간 인접 함수를 먼저 계산한다. 두 퍼지 클러스터 $\tilde{F}_p$, $\tilde{F}_q$와 주어진 소속 정도 $\mu$에 대해서 인접함수(proximity function) $f(\mu)$는 다음과 같이 정의된다:

$$f(\mu : \tilde{F}_p, \tilde{F}_q) = \sum_{j=1}^{n}\delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q)\omega(x_j) \quad (9)$$

$$\delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q) = \begin{cases} 1.0 & \text{if } \mu \leq MIN(\mu_{\tilde{F}_p}(x_j), \mu_{\tilde{F}_q}(x_j)) \\ 0.0 & \text{otherwise} \end{cases} \quad (10)$$

$\delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q)$는 데이터 $x_j$에 대해 소속정도 $\mu$에서 두 클러스터가 인접하는지를 결정한다. 두 클러스터 모두 $\mu$ 값 이상이면 인접성은 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 $\mu$에서 두 클러스터간의 인접도 계산을 도식화한 것이다. 전체 데이터 $x_j \in X$에 대해서 $x_j \in r$만이 식 10에 의해서 1.0의 인접도를 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.

*Definition 1* (클러스터 인접도) 패턴 행렬 $U$에 속하는 두 퍼지 클러스터 $\tilde{F}_p$와 $\tilde{F}_q$에 대해서, 각 소속 정도 $\mu$에서의 인접함수 $f(\mu : \tilde{F}_p, \tilde{F}_q)$가 주어진 경우, 두 클러스터간의 전체 인접도

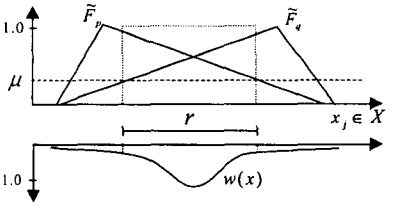


Fig. 2. 소속정도 $\mu$에서 두 클러스터간 인접도 $f(\mu)$

$S(\tilde{F}_p, \tilde{F}_q)$는 다음과 같이 정의된다

$$S(\tilde{F}_p, \tilde{F}_q) = \sum_{\mu} f(\mu : \tilde{F}_p, \tilde{F}_q) = \sum_{\mu} \sum_{j=1}^{n} \delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q)\omega(x_j) \quad (11)$$

$S(\tilde{F}_p, \tilde{F}_q)$는 $f(\mu : \tilde{F}_p, \tilde{F}_q)$를 전체 소속정도의 범위에 해서 계산한 것이다. 따라서, $S(\tilde{F}_p, \tilde{F}_q)$가 상대적으로 낮은 값을 가진다는 것은 클러스터 $\tilde{F}_p$와 클러스터 $\tilde{F}_q$의 인접도가 낮다는 것을 의미하므로, 두 클러스터가 잘 분할 되었다는 것을 알 수 있다. 두 클러스터간 인접도 정의를 기반으로 본논문에서 제안하는 퍼지 클러스터 평가 인덱스는 다음과 같이 기술할 수 있다.

*Definition 2* (제안하는 평가 인덱스) 두 퍼지 클러스터간의 인접도 $S(\tilde{F}_p, \tilde{F}_q)$와 분할된 클러스터의 수 $c$가 주어진 경우, 제안하는 퍼지 클러스터 평가 인덱스 $V_{proposed}(U, V : X)$는 다음과 같이 정의된다

$$V_{proposed}(U, V : X) = \frac{1}{{}_cC_2} \underbrace{\sum_{p=1}^{c} \sum_{q=1}^{c}}_{p \neq q} S(\tilde{F}_p, \tilde{F}_q) \quad (12)$$

여기서 ${}_cC_2$는 클러스터간 인접도 계산 수를 나타내므로, $V_{proposed}$는 분할에 속하는 모든 클러스터들 간의 평균 인접도를 나타내게 된다. 그러므로 $V_{proposed}$의 값이 낮을 수록 좋은 분할 결과라고 할 수 있다. 최적의 분할 결과 또는 최적의 클러스터 수는 $c = 2, 3, ..., c_{max}$에 대해서 $V_{proposed}$를 최소화 시킴으로써 계산할 수 있다.

## 4. 실험 및 분석

제안된 인덱스의 신뢰성과 효율성을 보이기 위해서, 다섯 가지의 표준 데이터 집합에 대해서 기존의 인덱스들과 성능 비교 실험을 수행하였다: $V_{PC}$ [2], $V_{PE}$ [3], $V_{XB}$ [5], $V_K$ [7], $V_{CWB}$ [8], $V_{B_{crit}}$ [9]. 각 실험 집합에 대해서, 표준 FCM 알고리즘이 산출한 분할 결과를 $c = 2, .., c_{max}$ 범위에서 평가하였다. 본 논문에서는 지면 관계상 대표적인 하나의 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 $V_{proposed}$가 정확히 8-클러스터로 분할한 경우가 최적이라고 계산하였다. 이와 달리, $V_{PC}$와 $V_{PE}$는 2개의 클러스터가 최적이라는 결과를, $V_{XB}$와 $V_K$는 최적 클러스터가 6이라고 제시했다. $V_{B_{crit}}$는 최적 분할로 5를 제시하였다.

TABLE I

STARFIELD에 대한 인덱스들의 평가 값 ($c = 2, .., c_{max} = 8$)

| $c$ | $V_{PC}$ | $V_{PE}$ | $V_{XB}$ | $V_K$ | $V_{CWB}$ | $V_{B_{crit}}$ | $V_{proposed}$ |
|---|---|---|---|---|---|---|---|
| 2 | **0.73** | **0.18** | 0.24 | 16.04 | 0.17 | 4.90 | 148.60 |
| 3 | 0.66 | 0.26 | 0.12 | 8.29 | 0.12 | 4.23 | 93.33 |
| 4 | 0.62 | 0.32 | 0.12 | 8.74 | 0.10 | 4.19 | 94.67 |
| 5 | 0.63 | 0.33 | 0.11 | 8.16 | 0.09 | **4.09** | 78.64 |
| 6 | 0.65 | 0.33 | **0.10** | **8.09** | 0.08 | 4.30 | 62.04 |
| 7 | 0.66 | 0.33 | 0.11 | 9.61 | 0.07 | 4.66 | 57.83 |
| 8 | 0.67 | 0.33 | 0.12 | 10.42 | **0.07** | 5.10 | **50.59** |

TABLE II

다섯 데이터 집합에 대한 각 인덱스들의 최적 클러스터 수

| Data | $V_{PC}$ | $V_{PE}$ | $V_{XB}$ | $V_K$ | $V_{CWB}$ | $V_{B_{crit}}$ | $V_{proposed}$ |
|---|---|---|---|---|---|---|---|
| D1 | **3** | 2 | **3** | **3** | 7 | **3** | **3** |
| D2 | 2 | 2 | 6 | 6 | **8** | 5 | **8** |
| D3 | **2** | **2** | **2** | **2** | 3 | 6 | **2** |
| D4 | **3** | **3** | **3** | 2 | **3** | **3** | **3** |
| D5 | **2** | **2** | **2** | **2** | **2** | **2** | **2** |

테이블 II는 다섯 데이터 집합에 대해서 각 평가 인덱스들이 계산한 최적 클러스터 수를 표시한 것이다: D1(BENSAID, $c = 3$) [10], D2(STARFIELD, $c = 8$) [5], D3(IRIS, $c = 2$) [4], D4(X30, $c = 3$) [11], D5(BUTTERFLY, $c = 2$) [7] (괄호안의 숫자는 해당 데이터 집합에서 알려진 최적 클러스터 수이다). 결과에서 보는 바와 같이 제안된 인덱스 $V_{proposed}$는 모든 데이터 집합에 대해서 정확한 평가 결과를 제시한다. 이에 반해, $V_{PC}$와 $V_{PE}$ 그리고 $V_{XB}$는 STARFIELD 데이터에서 최적의 $c$를 찾지 못한다. 더욱이, $V_{PE}$는 BENSAID 데이터에서도 실패하는 것을 알 수 있다. $V_K$는 STARFIELD와 X30 데이터에서, $V_{CWB}$는 BENSAID와 IRIS 데이터에서 최적의 클러스터 분할을 찾지 못한다. $V_{B_{crit}}$의 경우 STARTFIELD와 IRIS 데이터에 대해서 실패하는 결과를 보여준다.

## 5. 결론

본 논문에서는 새로운 퍼지 클러스터 평가 인덱스가 제안하였다. 제안된 인덱스는 기존 인덱스들의 한계점을 극복하기 위해서 클러스터간의 인접도를 이용하였다. 클러스터 인접도는 클러스터간의 중첩을 계산하는 척도로 사용되며, 인접도가 낮을 수록 잘 분리된 분할 결과를 제공된다. 따라서 최적의 분할 결과는 제안된 인덱스를 최소화시키는 방향으로 수렴된다. 실험 결과에서 살펴본 바와 같이 다양한 데이터 집합에서 기존 인덱스들보다 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] A.M. Bensaid, et al (1996) "Validity-guided (re)clustering with applications to image segmentation", IEEE Transactions on Fuzzy Systems, 4(2):112-123.

[11] 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.