새로운 클러스터 평가 지표

A Novel Cluster Validation Index

서석태*, 손세호**, 이인근***, 정혜천****, 권순학*
영남대학교 전기공학과

Suk. T. Seo, Seo. H. Son, In. G. Lee, Hye. C. Jeong, Soon. H. Kwon
* Dept. of Electrical Engineering, Yeungnam University
** Top Engineering Co., Ltd
*** Netblue Co., Ltd
**** Korea Textile Machinery Research Institute
E-mail : kenneth78@yumail.ac.kr

요 약

기존의 클러스터 평가 지표(cluster validation index)는 클러스터의 개수가 커질수록 클러스터 평가 지표 값이 단조 감소하는 경향을 보인다. 최근에 이러한 단점을 보완하는 새로운 클러스터 평가 지표가 본 논문 저자중의 하나에 의해 제안되었으나, over-clustering의 단점을 지니고 있다. 본 논문에서는, 클러스터 평가 지표 값이 단조 감소 및 over-clustering을 방지할 수 있는 새로운 클러스터 평가 지표를 제안하고, 여러 가지 예제를 통하여 새롭게 제안된 평가 지표의 타당성을 보인다.

Key words : 클러스터 평가 지표, 단조 감소, over-clustering

1. 서론

Zadeh의 퍼지 집합 이론이 제안된 이후에 제어, 패턴 인식, 그리고 클러스터링 등등의 수많은 분야에서 퍼지 집합을 기초로 한 연구가 행해지고 있다. 이러한 연구의 기본 아이디어는 적절한 방법에 의해서 언어변수를 기반으로 정의된 퍼지 규칙과 소속도 함수 값을 이용하여 주어진 시스템을 모델로 표현하는 것이다. 이러한 퍼지 모델링을 위하여 다양한 퍼지 클러스터링 방법들이 이용되고 있다[10].

클러스터 분석은 p-차원에서 n개의 데이터 집합 $X=\{x_1,\cdots,x_n\}\subset R^p$ 에 대하여 제시된 수의 그룹 혹은 클러스터로 각 원소들을 배치하는 것이다. 이와 같은 클러스터 분석을 위한 대표적 클러스터링 방법으로는 Bezdek의 Fuzzy c-mean (FCM) 알고리즘[2]과 Krishnappuram과 Keller의 가능성 기반 퍼지 클러스터링 알고리즘[3]을 들 수 있다.

Milligan의 논문[4]에서도 지적한바와 같이, 클러스터 분석은 클러스터링 방법뿐만 아니라 클러스터링 원소, 클러스터링 변수, 변수의 정규화, 관련성 측정, 클러스터 개수 설정 등등에 관한 분야를 포함하고 있다. 최근 들어, 많은 논문들이 클러스터 평가 지표, 즉 FCM 알고리즘과 같은 클러스터링 방법들에 의해서 생성된 데이터들의 분류 정당성에 관한 평가에 많은 관심을 보이고 있다[5-9]. Pal과 Bezdek의 분석[9]에 따르면 Fukuyama-Sugeno 지표[6]는 가중치 지수 m의 높고 낮은 값에 민감하고, 이러한 민감성 때문에 신뢰할 수가 없다고 한다. Xie-Beni 지표[8]는 클러스터 개수의 선택 뿐 아니라 가중치 지수 m의 선택에 있어서 1.01-7까지의 넓은 범위의 응답을 제공한다고 한다. 이러한 분석에 기초해서 그들은 [1.5 2.5]사이의 간격에서 이들의 평균값이자 중간값인 m=2의 가중치 지수를 최고의 값으로 제시하였고, 이 값은 FCM을 사용하는 사람들에 의해서 많이 사용되어지고 있다.

그럼에도 불구하고, Xie-Beni 지표는 클러스터의 개수 c가 커져 데이터의 개수 n에 가까워질수록 단조 감소하여 클러스터 타당성 평가 지표로서의 역할을 수행하지 못하게 되는 단점을 가지고 있다. Xie-Beni는 이러한 현상을 제거하기 위한 하나의 방법으로 적절한 문책함수(punishing function)를 사용하여 이러한 단조 감소 경향을 제거 할 수 있을 것이라고 제안하였으나, 함수를 어떻게 선택하는지에 대해서는 논의하지 않았다. 이러한 문제점을 극복하기 위하여 Kwon[10]이 Xie-Beni 지표에 문책함수를 추가한 형태의 클러스터 타당성 평가 지표를 제안하였으나 Over-clustering 이라는 단점을 지니고 있다.

본 논문에서는 기존의 클러스터 평가 지표의 단조 감소 경향의 문제점을 극복하고 또한 Kwon의 지표가 갖는 Over-clustering 특성을 개선하는 새로운 클러스터 평가 지표를 제안한다.

2. FCM 알고리즘, 클러스터 평가 지표

FCM 알고리즘은 멤버십 함수 값 u_{ij}와 클러스터 센터 값 v_{ij}를 가지는 아래의 함수 값을 최적화 하는 문제이다.

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

$U=[u_{ij}]$ 는 c x n 행렬, c는 클러스터 개수, n은 데이터의 개수이며, 다음의 조건을 만족한다.

$$M_{fcm}=\{U\in R^{cn} | u_{ij}\in[0,1]\ \forall i,j; 0<\sum_{j=1}^{n}u_{ij} < n\ \forall i, \text{and} \sum u_{ij}=1\ \forall j\ \} \quad (2)$$

$V=(v_1,\cdots,v_c)$ 는 클러스터 중심 벡터, $v_i\in R^p$, $c\geq i\geq 1$ 이며, $\| \cdot \|$ 는 내적이다. X의 최적 분할 U^*는 다음과 같은 필수 조건식의 반복에 의해 구해진 값들 중 최소 J_m의 값을 가지는 (U^*,V^*)의 쌍으로부터 구해진다.

Fuzzy c-means[2]: If

$D_{ij}=\| x_j-v_i \|_A>0\ \forall i,j$

가중치 지수 m>1,데이터 집합 X는 c<n, 따라서 $(U,V)\in M_{fcm}\times R^{cp}$ 는 다음과 같을 때 최소 J_m을 가진다.

$$u_{ij}=\left[\sum_{k=1}^{c}\left(\frac{D_{ijA}}{D_{jiA}}\right)^{\frac{2}{m-1}}\right]^{-1},1\leq i\leq c;1\leq j\leq n$$

$$v_i=\frac{\sum_{j=1}^{c}u_{ij}^m x_j}{\sum_{j=1}^{n}u_{ij}^m},1\leq i\leq c \quad (3)$$

만약 $D_{ijA}=0$ 이라는 특이성을 가지는 i와 j의 경우, $D_{ijA}>0$ 인 각각의 u_{ij}에 0을 할당하고, $D_{ijA}=0$ 의 값을 가지는 x_k에 대해서 임의의 멤버십 함수 값을 배당한다. 식(3)의 몇몇 제한적 특성에 관한 연구는 Pal과 Bezdek[9]에 의해서 연구되어졌으므로 여기서는 언급하지 않는다.

2.1 기존의 클러스터 평가 지표

Dunn's normalized partition entropy[5]:

$$V_D(U)=\frac{n}{n-c}v_{PE}=-\frac{1}{n-c}\sum\sum u \log(u) \quad (4)$$

Bezdek's partition coefficient[2]:

$$v_{pc}(U)=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^2 \quad (5)$$

Bezdek's partition entropy[2]:

$$v_{PE}(U)=-\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\log_a(u_{ij})$$

log의 지수$a\in(1,\infty)$, $u_{ij}=0$ 일 때 $u_{ij}\log(u_{ij})\cong 0$

Fukuyama-Sugeno Index[6]:

$$V_{FS}(U,V;X)=\sum\sum u \cdot \| x-v \| - \| v-\bar{v} \| \quad (7)$$

Xie-Beni Index[8]:

$$v_{XB}(U,V;X)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^2 \| x_j-v_i \|^2}{n\left[\min_{i\neq k} \| v_i-v_k \|^2\right]} \quad (8)$$

Extended FCM Xie-Beni Index[8]:

$$v_{XB}(U,V;X)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^m \| x_j-v_i \|^2}{n\left[\min_{i\neq k} \| v_i-v_k \|^2\right]} \quad (9)$$

식(8)번의 Xie-Beni Index v_{XB}가 가중치 m과 클러스터의 개수에 있어서 가장 넓은 범위에 대해서 가장 좋은 응답을 제공하기 때문에 본 논문에서는 식(8)에 대해서 고찰한다.

v_{XB}는 클러스터의 개수가 데이터의 개수 n에 가까워질수록 단조감소 하는 경향을 가지고 있고 이러한 단조 감소의 영향에 의한 오류를 피하기 위해서 Xie-Beni는 c에 따라 v_{XB}를 도시하였다. 그들은 그래프를 통하여 최대 클러스터 개수에서 단조 감소 시작점을 찾았고, v_{XB}를 최소화 하는 c의 값을 선택하였다. 이러한 방법은 적절한 c

의 값을 제공하였지만, 너무 번거로움이 있기 때문에 효과적이지 않다.

Kwon Index[10]:

$$v_K(U,V;X)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^2\|x_j-v_i\|^2+\frac{1}{c}\sum_{i=1}^{c}\|v_i-\bar{v}\|^2}{\min_{i\neq k}(\|v_i-v_k\|^2)} \quad (10)$$

여기서, $\bar{v}=\frac{1}{n}\sum_{j=1}^{n}x_j$

Kwon의 지표는 문책함수의 과다한 책정으로 Over-clustering 현상을 나타낸다.

2.2 Kwon Index의 확장

일반적으로, 클러스터링은 클러스터 내부 유사성 및 클러스터간 비유사성을 최대화하는 전략을 기반으로 한다. 이러한 배경 하에 본 논문에서는 새로운 클러스터 평가 지표 v_{Ke}를 제안한다.

$$v_{Ke}(U,V;X)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^m\|x_j-v_i\|^2+\frac{1}{n}\sum_{i=1}^{c}\|v_i-\bar{v}\|^2}{\min_{i\neq k}(\|v_i-v_k\|^2)} \quad (11)$$

식(11)의 분자의 첫 번째 항은 내부 클래스의 유사성을 측정하며, 이 값이 작을수록 클래스는 더욱더 유사하다 할 수 있다. 두 번째 항은 클러스터 개수 c가 데이터의 개수에 가까워질수록 단조 감소하는 경향을 제거하는 문책함수이다. 분모 항은 클러스터 중심 사이의 최소거리를 나타낸다. 이 값이 클수록 클러스터 사이에는 보다 큰 비유사성이 존재함을 의미한다.

클러스터 지표의 극한 작용을 조사한다.

Xie-Beni Index, $c\to n$:

$$\lim_{c\to n}\|x_j-v_j\|^2=0$$

$$\lim_{c\to n}v_{XB}(U,V;X)=\lim_{c\to n}\frac{\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^2\|x_j-v_j\|^2}{n\left[\min_{i\neq k}(\|v_i-v_k\|)^2\right]}=0 \quad (12)$$

식(12)로부터, 우리는 Xie-Beni Index가 c의 값이 커짐에 따라서 FCM으로부터 얻어진 (U,V)에 대한 평가 기능을 상실함을 알 수 있다.

제안된 평가 지표는, 식(11)에서처럼 $c\to n$으로 수렴하는 경우,

$$\lim_{c\to k}v_{Ke}(U,V;X)=\lim_{c\to k}\frac{\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^m\|x_j-v_i\|^2+\frac{1}{n}\sum_{i=1}^{c}\|v_i-\bar{v}\|^2}{\min_{i\neq k}(\|v_i-v_k\|^2)}$$

$$=\frac{\frac{1}{n}C_X}{\min_{i\neq k}(\|v_i-v_k\|)^2} \quad (13)$$

여기서, C_X는 X의 전체 분산 행렬이다. 식(13)으로부터 제안된 평가지표가 c의 값이 커짐에 따라 FCM으로부터 얻어진 (U,V)에 대한 평가 기능을 보존하고 있음을 확인할 수 있다.

3. 클러스터 평가 지표에 관한 예제

	Data X_1 (c=2)		Data X_2 (c=3)		Data X_3 (c=3)	
Points	x	y	x	y	x	y
1	0	0	-3	0	1.5	2.5
2	0	2	-2	-1	1.7	2.6
3	0	4	-2	0	1.2	2.2
4	1	1	-2	1	2	2
5	1	2	-1	0	1.7	2.1
6	1	3	-1	3	1.3	2.5
7	2	2	0	1	2.1	2
8	3	2	0	2	2.3	1.9
9	4	2	0	3	2	2.5
10	5	1	0	4	1.9	1.9
11	5	2	1	0	5	6.2
12	5	3	1	3	5.5	6
13	6	0	2	-1	4.9	5.9
14	6	2	2	0	5.3	6.3
15	6	4	2	1	4.9	6
16			3	0	5.8	6
17					5.5	5.9
18					5.2	6.1
19					6.2	6.2
20					5.6	6.1
21					10.1	12.5
22					11.2	11.5
23					10.5	10.9
24					12.2	12.3
25					10.5	11.5
26					11	14
27					12.2	12.2
28					10.2	10.9
29					11.9	12.7
30					12.9	12

표 1. 예제 데이터 집합 X_1, X_2, X_3

예제 1: 그림 1의 2차원에서 15개의 데이터를 가지고 있는 데이터 집합을 고찰한다.

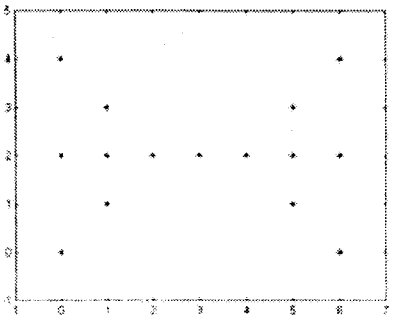

그림 1. 예제 1 : X_1

예제 2: 그림 2에 나타나있는 2차원공간에서 16개의 데이터를 가지고 있는 데이터 집합을 고찰한다.

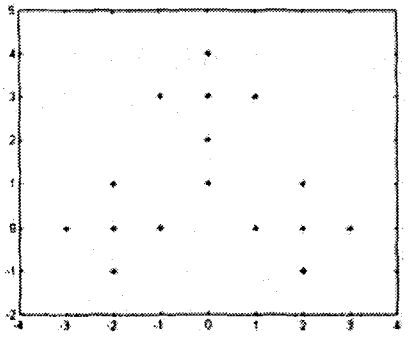

그림 2. 예제 2 : X_2

예제 3: 그림 3과 같은 2차원에서 30개의 데이터를 가지고 있는 집합을 고찰한다.

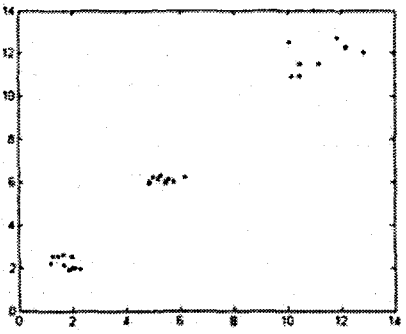

그림 3. 예제 3 : $X30$

m	X_1 : c^*=2		X_2 : c^*=3		$X30$: c^*=3	
	v_{XB}	v_{Ke}	v_{XB}	v_{Ke}	v_{XB}	v_{Ke}
1.5	14	2	15	3	29	3
2.0	14	2	15	3	29	3
3.0	14	2	15	3	29	3
4.0	14	2	15	3	29	3
5.0	14	2	15	3	29	3
6.0	14	2	15	3	29	3
7.0	14	2	15	3	29	3

표 2. X_1,X_2,$X30$의 m값의 변화에 따른 c개수

표 2로부터 우리는 제안된 클러스터 평가 지표가 기존의 클러스터 평가 지표에서 보이던 단조 감소 현상을 제거하는 것을 확인 하였다. 실험에 사용된 $\varepsilon = 1e-9$ 이며, m=1.5, 2, 3, 4, 5, 6, 7이다.

4. 결론

본 논문에서, 클러스터의 개수가 데이터의 개수에 가까워질수록 기존의 클러스터 평가 지표가 가지는 단조 감소하는 경향 및 Over-clustering 현상을 제거하는 새로운 클러스터 평가 지표를 제안하였으며, 여러 가지 예제를 통하여 제안된 지표의 타당성을 확인하였다.

5. 참고문헌

[1] M. Sugeno and T. Yasukawa, "A fuzzy logic based approach to qualitative modeling," IEEE Trans. Fuzzy Syst., Vol.1, No. 1, pp. 7-31, 1993.

[2] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Pleum, New York, 1981.

[3] R. Krishnappuram and J. M. Keller,"A possibilistic approach to clustering," IEEE Trans. Fuzzy Syst., Vol. 1, No. 2, pp. 98-110, 1993.

[4] G. W. Milligan, "Clustering validation," in Clustering and Classification, P. Arabie, L. J. Hubert and G. D. Soete, ED. World Scientific, Singapore, 1996.

[5] J. C. Dunn, "Indices of partition fuzziness and the detection of clusters in large data sets," in Fuzzy Automata and Decision Processes, M. M. Gupta, Ed. Elsvier, New York, 1976.

[6] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," in Proc. 5th Fuzzy Syst. Symp., pp. 247-250, 1989.

[7] J. C. Bezdek and N. K. Pal, "Some new indexes of cluster validity," IEEE Trans. Systems, Man, and Cyber-Part B, Vol. 29, No. 3, pp. 301-315, 1998.

[8] X. L. Xie and G. A. Beni,"Balidity measure for fuzzy clustering," IEEE Trans. Pattern and Machine Intell., Vol. 3, No. 8, pp. 841-846, 1991.

[9] N. K. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," IEEE Trans. Fuzzy Syst., Vol. 3, No. 3, pp. 370-379, 1995.

[10] S. H. Kwon, "Cluster validity index for fuzzy clustering," Electronics Letters, Vol. 34, No. 22, pp. 2176-2177, 1998.