A Cluster Validity Index for Fuzzy Clustering

퍼지 클러스터링의 타당성 평가 기준

Soon H. Kwon
권 순 학

School of Electrical and Electronic Eng., Yeungnam University
영남대학교 전기전자공학부

요 약

본 논문에서는, 퍼지 클러스터의 수가 증가함에 따라 나타나는 퍼지 클러스터링 타당성 평가 기준의 단조 감소 현상을 억제하는 새로운 퍼지 클러스터링 타당성 평가 기준을 제시한다. 또한, 제시된 평가 기준의 성질을 조사하고 기존의 퍼지 클러스터링 타당성 평가 기준과의 차이점에 대하여 논한다. 마지막으로, 퍼지 클러스터링에 자주 인용되는 몇 가지 전형적인 자료에 대한 모의 실험을 통하여 제시된 평가 기준의 효용성을 보인다.

Abstract

In this paper, a new cluster validation index which is heuristic but able to eliminate the monotonically decreasing tendency occurring in which the number of cluster c gets very large and close to the number of data points n is presented. We review the FCM algorithm and some conventional cluster validity criteria, discuss on the limiting behavior of the proposed validity index, and provide some numerical examples showing the effectiveness of the proposed cluster validity index.

I. Introduction

Since Zadeh's formulation of fuzzy set theory, many fuzzy set-based approaches to fields such as control, pattern recognition, decision-making, and clustering have been developed and applied to systems with uncertainty. The basic idea of these approaches is to represent the uncertainty of the given systems by means of fuzzy rules and their membership functions defined over appropriate discourses. One of the most prominent applications of it may be a fuzzy logic-based modeling by means of fuzzy clustering [10].

Cluster analysis is to place elements into groups or clusters suggested by a given data set $X=\{x_1, \ldots, x_n\} \subset R^p$ which are n points in the p-dimensional space for summarizing data or finding "natural" or "real" substructures in the data set. The Fuzzy C-Means (FCM) algorithm [1] and its derivatives based on the possibilistic approach [3,4] for the cluster analysis have been the dominant approaches in both theory and practical applications of fuzzy techniques to unsupervised classification for the last two decades.

As pointed out by Milligan [2], a cluster analysis will not only refer to clustering methods such as the FCM and the possibilistic approach but also to the overall sequence of steps such as clustering elements, clustering variables, variable standardization, measure of association, number of clusters, interpretation, testing, and replication. In recent years, many literatures have paid a great deal of attention to cluster validity issues, and many functionals have been proposed for validation of partitions of data produced by the FCM algorithm

[5,6,7,8,9,11]. According to the Pal and Bezdek's analysis [9], the Fukuyama-Sugeno index [6] is sensitive to both high and low values of the weighting exponent m and may be unreliable because of this. The Xie-Beni index provided the best response over a wide range of choices for the number of clusters, (2-10), and for the weighting exponent m from 1.01-7. On the basis of their analysis, they suggested that the best choice for the weighting exponent m may be probably in the interval [1.5, 2.5], whose mean and midpoint, m=2, have often been the preferred choice for many users of the FCM.

However, the Xie-Beni index v_{XB} has a flaw that is monotonically decreasing when the number of cluster c gets very large and close to the number of data points n. Xie and Beni suggested that an ad hoc punishing function should be imposed to eliminate the monotonically decreasing tendency, but not discussed how to choose the function. It is highly recommended to impose a punishing function, which is a function of the number of cluster, to eliminate the monotonically decreasing tendency of the cluster validation indexes as like as the statistical model selection criteria do.

The author has proposed a new cluster validity index for fuzzy clustering [11]. In this paper, the cluster validation index which is heuristic but able to eliminate the decreasing tendency occurring in which the number of cluster c gets very large and close to the number of data points n is presented. The FCM algorithm and some conventional cluster validity criteria are presented and the limiting behavior of the proposed validity index will be discussed.

This paper is organized as follows. In section 2, we review the FCM algorithm and some cluster validity criteria, and present a new cluster validity index. Section 3 describes some numerical examples showing the effectiveness of the presented cluster validity measure.

II. Fuzzy C-Means Algorithm and Cluster Validity

The Fuzzy C-Means (FCM) algorithm is a constrained optimization problem which minimizes the following objective function with respect to membership functions u_{ij} and cluster centroid v_i,

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

where $U=[u_{ij}]$ is a c x n matrix, c is the number of clusters, n is the number of data points, satisfying the conditions in (2),

$$M_{fcn} = \left\{ U \in R^{cn} \,\middle|\, u_{ij} \in [0,1] \; \forall i,j; \;\; 0 < \sum_{j=1}^{n} u_{ij} < n \;\; \forall i, \; and \; \sum_{i=1}^{c} u_{ij} = 1 \; \forall j \right\} \tag{2}$$

$V=(v_1, \ldots, v_c)$ is a vector of cluster centers, $v_i \in R^p$ for $c \geq i \geq 1$ and $\|\bullet\|$ denotes any inner product norm. Optimal partitions U^* of X are taken from pairs (U^*, V^*) that are local minimizers of J_m obtained by iteration through the following necessary conditions.

Fuzzy C-Means Theorem [1]:

If

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

the weighting exponent m>1, and a data set X contains c < n distinct points, then $(U,V) \in M_{fcn} \times R^{cp}$ may minimize J_m only if

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{D_{ijA}}{D_{jkA}} \right)^{\frac{2}{m-1}} \right]^{-1}, \;\; 1 \leq i \leq c; \;\; 1 \leq j \leq n \tag{3}$$

$$v_i = \frac{\sum_{j=1}^{c} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}}, \quad 1 \le i \le c.$$

If for some i and j, $D_{ijA} = 0$, a singularity occurs, then assign 0's to each u_{ij} for which $D_{ijA} > 0$, and distribute membership functions arbitrary across the x_k's for which $D_{ijA} = 0$, subject to the constraints in (2). Some limiting properties of (3) have been studied by Pal and Bezdek [9] and are not discussed here.

Among a class of cluster validity functionals such as
the Dunn's normalized partition entropy [5] :

$$v_D(U) = \frac{n}{n-c} v_{PE} = -\frac{1}{n-c} \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij} \log_a(u_{ij}), \tag{4}$$

the Bezdek's partition coefficient [1] :

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

the Bezdek's partition entropy [1] :

$$v_{PE}(U) = -\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij} \log_a(u_{ij}), \tag{6}$$

where logarithmic base a $\in (1,\infty)$ and $u_{ij} \log(u_{ij}) \cong$ whenever $u_{ij} = 0$,
the Fukuyama-Sugeno index [6] :

$$v_{FS}(U,V;X) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^{m} \left(\|x_j - v_i\|^2 - \|v_i - \bar{v}\|^2 \right), \tag{7}$$

the Xie-Beni index [8]:

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

and the extended FCM Xie-Beni index [8] :

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

We consider only the Xie-Beni index v_{XB} given by (8) because it provides the best response over a wide range of choices for the number of clusters and for the weighting exponent m as discussed in the introduction.

Xie and Beni stated that v_{XB} decreases monotonically when the number of clusters c is close to n. To avoid the indetermination due to the monotonicity, they recommended plotting v_{XB} as a function of c, finding the starting point of the monotonic epoch as the maximum cluster number to be considered, and then selecting a value c minimizing v_{XB}. Because it requires a cumbersome procedure to find an optimum value of c, it may not be a sufficiently good cluster validity index even though it provides good responses through the cumbersome procedures.

In clustering, it attempts to maximize intra-class similarity and inter-class differences. In this sense, a new cluster validity index v_K is defined as

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

where $\bar{v} = \frac{1}{n} \sum_{j=1}^{n} x_j$.

The first term of the numerator in (10) measures the intra-class similarity, that is, how compact each and every class is. The more similar (compact) the classes, the smaller it is. It is independent of the number of data points. The second term of the numerator in (10) is an ad hoc punishing function imposed to eliminate the decreasing tendency occurring when the number of cluster c gets very large and close to the number of data points n. The denominator in (10) which is the minimum distance between cluster centroids measures the inter-class difference. A larger value of it indicates that every cluster is well-seperated. Our goal is to find the fuzzy c-partition with the smallest value of v_K.

In order to investigate the limiting behavior of the proposed index, we take a limit of the validity functional as c approaches n.

Xie-Beni index, c → n: Since

$$\lim_{c\to n}\|x_j - v_i\|^2 = 0, \tag{11}$$

we have

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

From (12), we can see that the Xie-Beni index loses its ability to validate (U,V) pairs from the FCM for the large value of c.

The proposed index, c → n: Since (11) holds for this case, we have

$$\begin{aligned}\lim_{c\to n} v_K(U,V;X) &= \lim_{c\to n}\frac{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{2}\|x_j - v_i\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|v_i - \bar{v}\|^2}{\min_{i\neq k}(\|v_i - v_k\|^2)}\\ &= \frac{\frac{1}{n}C_X}{\min_{i\neq k}(\|v_i - v_k\|^2)}\end{aligned} \tag{13}$$

where C_X is the total scatter matrix of X. From (13), we can see that the the proposed index keeps its ability to validate (U,V) pairs from the FCM for large value of c. Here, we do not discuss on the intuitive meaning and mathematical justification of the proposed index, which are required for the any new validity functional, because the research on those is on the way.

III. Numerical Examples On the Cluster Validity Index

In this section, we consider three examples of data sets to show the effectiveness of the proposed cluster validity. We first present a simple example with c=2 as the preferred clusters, which is known as the butterfly data set [1] to provide insights into the limiting behavior of the cluster validity indexes. We then present two examples which are the derivatives of the butterfly data set and have c=3 and c=4 as the preferred clusters, respectively.

Example 1: We consider the butterfly data set X_1 of 15 data points in p=2 dimensions shown in Fig.1. Data points (2,2), (3,2), and (4,2) form a bridge or neck between the wings of the butterfly. Another interpretation of the pattern is that points in the wings were drawn from two fairly distinct classes; points in the neck are noise.

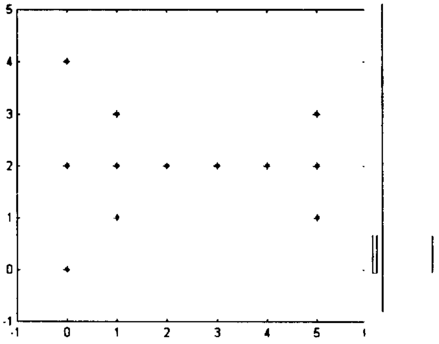

Fig. 1. Example 1 : X_1.

Example 2: We consider a set X_2 of 22 data points in p=2 dimensions shown in Fig.2. A data point (0,1) forms a bridge among three diamonds. Another interpretation of the pattern is that points in the each diamond were drawn from three fairly distinct classes; a point in the neck is noise.

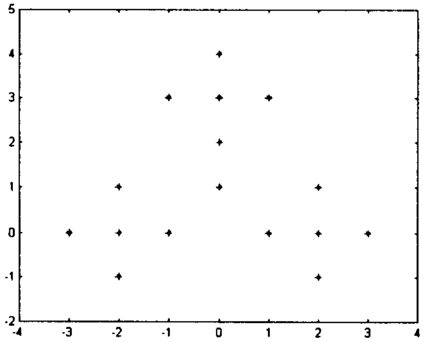

Fig. 2. Example 2 : X_2.

Example 3: We consider a the butterfly data set X_3 of 29 data points in p=2 dimensions shown in Fig.3. Data points in each triangular were drawn from four fairly distinct classes.

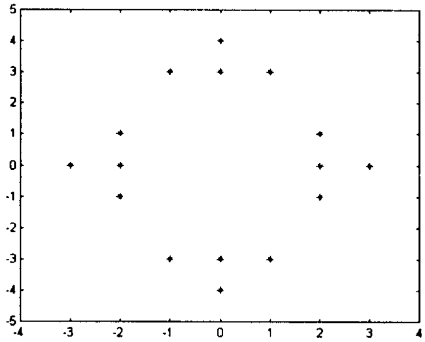

Fig. 3. Example 3 : X_3.

For each of data sets shown in the above we performed the FCM with the terminating criterion $\varepsilon = 1.0e\text{-}8 \geq$

$||U_t - U_{t-1}||$ for different weighting exponents m=1.2, 2.0, 3.0, 4.0, 5.0, 6.0 and 7.0, and c=2, 3, . . . , n-1. Table 1 shows index values by each of the Xie-Beni index and the proposed index on the data set X_1 for c=2 to 14, and m=1.2, 2.0, and 7.0.

Table 1. Index values on the data set X_1 for c=2 to 14, and m=1.2, 2.0, 7.0

c	The Xie and Beni index v_{XB}			The proposed index v_K		
	m=1.2	m=2.0	m=7.0	m=1.2	m=2.0	m=7.0
2	0.1059	0.0954	0.2242	2.5402*	2.3864*	4.4367*
3	0.2650	0.2054	0.5460	6.7261	6.0076	12.0636
4	0.2388	0.3972	1.6116	6.3251	15.6559	40.8956
5	0.1509	0.1061	1.3306	6.0418	5.3539	34.3596
6	0.2613	0.0946	1.0503	12.446	6.8607	36.0907
7	0.2289	0.0649	0.8068	10.4327	6.2112	27.5306
8	0.4220	0.1123	0.6485	27.4192	13.4601	25.4776
9	0.2307	0.1434	0.5180	18.9677	22.2546	26.6591
10	0.2512	0.1017	0.3827	27.5783	16.6737	23.8405
11	0.1874	0.0913	0.2997	22.3963	20.8977	21.3138
12	0.1216	0.0362	0.2182	20.4382	16.0205	22.6899
13	0.2048	0.0380	0.1020	47.5482	19.9721	21.2986
14	0.0332*	0.0128*	0.0450*	21.2304	19.1730	21.6036

Asterisks in Table 1 indicate the minimum index values obtained by each index for the weighting exponent m=1.2, 2.0, 7.0. Since the preferred value of c is 2, we see that the proposed index correctly points to the preferred value of c for each weighting exponent, but the Xie-Beni index points to c=14. This behavior is consistent with the fact, which is the Xie-Beni index loses its ability to validate (U,V) pairs from the FCM for the large value of c, discussed in the previous section. Table 2 lists the value of the number of clusters chosen by each of the Xie-Beni index and the proposed index.

Table 2. Values of c chosen by each index for the data sets X_2 and X_3

m	$X_2 : c^*=3$		$X_3 : c^*=4$	
	v_{XB}	v_K	v_{XB}	v_K
1.2	15	3	15	4

2.0	15	3	15	4
3.0	15	3	15	4
4.0	15	3	15	4
5.0	15	3	15	4
6.0	15	3	15	4
7.0	15	3	15	4

Since the preferred values of c are 3 and 4, respectively, we see that the proposed index correctly points to the preferred values c=3 and c=4 for each weighting exponent but the Xie-Beni index points to c=15 in every case. From these results, we conclude that the proposed cluster validity index shows the superior performance to the Xie-Beni index, and the Xie-Beni index may be unreliable.

IV. Conclusions

In this paper, we have presented a cluster validation index to eliminate the monotonically decreasing tendency, which is the typical flaw of the conventional cluster validity indexes, when the number of cluster gets very large and close to the number of data points. We have reviewed the FCM algorithm and some cluster validity criteria, and discussed on the limiting behavior of the presented validity index. Finally, numerical examples showing the effectiveness of the proposed cluster validity index have been provided.

Researches on the description of intuitive meaning, the mathematical justification, and applications to real data sets are on the way.

References

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

[2] G. W. Milligan, "Clustering Validation," in Clustering and Classification, P. Arabie, L. J. Hubert and G. D. Soete, Ed. World Scientific, Singapore, 1996.

[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] N. R. Pal, K. Pal and J. C. Bezdek, "A Mixed c-Means Clustering Model," in Proc. FUZZ-IEEE'97, 1997, pp. 11-21.

[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. Elsevier, 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., 1989, pp. 247-250 (in Japanese).

[7] J. C. Bezdek and N. K. Pal, "Some New Indexes of Cluster Validity," IEEE Trans. Systems, Man, and Cyber.-Part B, vol.28, no. 3, pp.301-315, 1998.

[8] X. L. Xie and G. A. Beni, "Validity 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] 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.

[11] Soon H. Kwon, "Cluster validity index for fuzzy clustering," Electronics Letters, 1998 (to be published).