LETTER

New Inter-Cluster Proximity Index for Fuzzy *c*-Means Clustering*

Fan LI[†a)], *Nonmember*, **Shijin DAI**[††], *Student Member*, **Qihe LIU**[†], *and* **Guowei YANG**[†], *Nonmembers*

SUMMARY This letter presents a new inter-cluster proximity index for fuzzy partitions obtained from the fuzzy *c*-means algorithm. It is defined as the average proximity of all possible pairs of clusters. The proximity of each pair of clusters is determined by the overlap and the separation of the two clusters. The former is quantified by using concepts of Fuzzy Rough sets theory and the latter by computing the distance between cluster centroids. Experimental results indicate the efficiency of the proposed index.

key words: *fuzzy c-means algorithm, Fuzzy Rough sets, inter-cluster proximity, cluster validity, fuzzy clustering*

1. Introduction

The fuzzy *c*-means (FCM) algorithm [1] is the dominant method for fuzzy clustering. The aim of FCM is to partition a given set of data points (patterns) $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subset R^p$ into c clusters represented as fuzzy sets $F_1, F_2, \cdots, F_c$. In FCM, a fuzzy partition is denoted as $(\mathbf{U}, V)$. $\mathbf{U} = [u_{ij}]$ is called the partition matrix, where u_{ij} is the membership value of x_j belonging to F_i satisfying $\sum_{i=1}^{c} u_{ij} = 1$ $(j = 1, 2, \cdots, n)$ and $0 < \sum_{j=1}^{n} u_{ij} < n$ $(i = 1, 2, \cdots, c)$. $V = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_c\}$ is the set of centroids of the c clusters. Obviously, $F_i(x_j) = u_{ij}$. The FCM objective function has the form of

$$J_m(\mathbf{U}, V) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \parallel \mathbf{x}_j - \mathbf{v}_i \parallel^2, \tag{1}$$

where $\|\cdot\|$ is a certain distance function and the exponent $m > 1$ is called a fuzzifier. FCM iteratively updates $\mathbf{U}$ and V to minimize $J_m(\mathbf{U}, V)$ until a certain termination criterion has been satisfied.

In FCM, if c is not known a priori, a cluster validity index must be used to evaluate qualities of fuzzy partitions for different values of c to find out the optimal cluster number. In general, a validity index is a composition of an intra-cluster similarity index and an inter-cluster proximity index. In most cited validity indices, e.g. the Xie-Beni index [2] and the Fukuyama-Sugeno index [3], the inter-cluster proximity of a fuzzy partition is considered as the cluster separation strength and estimated by using the distance between cluster centroids (or its variations). But this kind of indices is not effective for measuring the proximity caused by the overlap between clusters (see [4], [5]). To overcome this shortcoming, another kind of index has been proposed in recent years [4], [5], in which the proximity of two clusters involves only membership values of each data point belonging to the two clusters whereas the distance between cluster centroids is not taken into account.

Combining basic ideas of the above two kinds of indices, in this letter we propose a new index to evaluate the inter-cluster proximity of a fuzzy partition. The proximity of two clusters consists of the overlap and separation measures, which are quantified by using concepts of Fuzzy Rough sets theory and the cluster centroids distance, respectively. The inter-cluster proximity of a fuzzy partition is the average proximity of all possible pairs of clusters. Experimental results indicate the efficiency of the proposed index.

2. Basic Notions

Let $\mathfrak{U}$ be a nonempty finite set of objects and $\mathfrak{R}$ an equivalence relation on $\mathfrak{U}$. Let $[x]_{\mathfrak{R}}$ denote the equivalence class of x. The $\mathfrak{R}$-lower and $\mathfrak{R}$-upper approximations of a set $X \subseteq \mathfrak{U}$ are defined as follows [6]:

$$\underline{\mathfrak{R}}(X) = \{x \in \mathfrak{U} : [x]_{\mathfrak{R}} \subseteq X\}, \tag{2}$$

$$\overline{\mathfrak{R}}(X) = \{x \in \mathfrak{U} : [x]_{\mathfrak{R}} \cap X \neq \emptyset\}. \tag{3}$$

The approximation accuracy of the set X can be characterized numerically by the following coefficient [6]:

$$\alpha_{\mathfrak{R}}(X) = \frac{|\underline{\mathfrak{R}}(X)|}{|\overline{\mathfrak{R}}(X)|}. \tag{4}$$

In [7], Dubois and Prade proposed Fuzzy Rough sets theory, which extends concepts of the classical Rough sets theory [6] to fuzzy information systems. Then, their definitions were generalized in [8].

Let U be a nonempty set of objects. A fuzzy binary relation R on U is called a T-similarity relation if R satisfies:
(1) Reflexivity: $R(x, x) = 1, \forall x \in U$.
(2) Symmetry: $R(x, y) = R(y, x), \forall x, y \in U$.
(3) T-Transitivity: $R(x, z) \geq T(R(x, y), R(y, z)), \forall\, x, y, z \in U$, where T is a triangular norm.

Manuscript received July 10, 2007.
Manuscript revised September 25, 2007.

[†]The authors are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China.
[††]The author is with the School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China.

*This letter is supported by Hi-Tech research and development program of China (2005AA114030).

a) E-mail: lifan@uestc.edu.cn

DOI: 10.1093/ietisy/e91–d.2.363

Let F be a fuzzy subset of U and R a T-similarity relation, where T is a lower semi-continuous triangular norm. The R-lower and R-upper approximations of F, denoted by two fuzzy sets $\underline{R}(F)$ and $\overline{R}(F)$ respectively, are defined as [8]:

$$\underline{R}(F)(x) = \inf_{y \in U} I_T\{R(x,y), F(y)\}, \tag{5}$$

$$\overline{R}(F)(x) = \sup_{y \in U} T\{R(x,y), F(y)\}, \tag{6}$$

where I_T is the residuation implication of T, i.e. $I_T(a,b) = sup\{c \in [0,1] : T(a,c) \leq b\}$ for every $a, b \in [0,1]$.

3. Proposed Inter-Cluster Proximity Index

3.1 Motivations

An inter-cluster proximity index should indicate two kinds of information: the overlap and the separation between clusters. The cluster centroids distance can measure them to some extent, but is not sufficient. Let's consider two fuzzy partitions $(\mathbf{U}^{(a)}, V^{(a)})$ and $(\mathbf{U}^{(b)}, V^{(b)})$, each of which contains only two clusters, and cluster centroids distances are equal, i.e. $\| \mathbf{v}_1^{(a)} - \mathbf{v}_2^{(a)} \| = \| \mathbf{v}_1^{(b)} - \mathbf{v}_2^{(b)} \|$. Obviously, by using the cluster centroids distance, the inter-cluster proximity of $(\mathbf{U}^{(a)}, V^{(a)})$ equals to that of $(\mathbf{U}^{(b)}, V^{(b)})$, regardless of the possible difference in cluster overlap of the two partitions. In our opinion, since the cluster centroids distance provides useful but insufficient information about the inter-cluster proximity, combining it with the overlap measure may provide preferable results.

3.2 Detailed Descriptions

In general, the distance between two data points can qualify their similarity, i.e. the longer distance between them, the less degree they being "similar", and vice versa. Thus, we have the following definition.

Definition 1: Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subset R^p$ be a given set of data points. A fuzzy binary relation S on X is defined as: $\forall \mathbf{x}_i, \mathbf{x}_j \in X$,

$$S(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\| \mathbf{x}_i - \mathbf{x}_j \|}{d_{max}} \tag{7}$$

where $d_{max} = max_{i,j}\{\| \mathbf{x}_i - \mathbf{x}_j \|\}$.

Proposition 1: S is a T_L-similarity relation, where T_L is the Lukasiewicz t-norm: $T_L(a,b) = \max\{0, a+b-1\}$ for every $a, b \in [0,1]$.

Proof. Reflexivity and Symmetry are obvious. We prove T_L-transitivity as follows. $\forall \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in X$, we have: $T_L(S(\mathbf{x}_i, \mathbf{x}_j), S(\mathbf{x}_j, \mathbf{x}_k)) = max\{0, 1 - \frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{d_{max}} + 1 - \frac{\|\mathbf{x}_j - \mathbf{x}_k\|}{d_{max}} - 1\} = max\{0, 1 - \frac{\|\mathbf{x}_i - \mathbf{x}_j\| + \|\mathbf{x}_j - \mathbf{x}_k\|}{d_{max}}\}$. By the triangle inequality, $\| \mathbf{x}_i - \mathbf{x}_j \| + \| \mathbf{x}_j - \mathbf{x}_k \| \geq \| \mathbf{x}_i - \mathbf{x}_k \|$. So $T_L(S(\mathbf{x}_i, \mathbf{x}_j), S(\mathbf{x}_j, \mathbf{x}_k)) \leq max\{0, 1 - \frac{\|\mathbf{x}_i - \mathbf{x}_k\|}{d_{max}}\} = S(\mathbf{x}_i, \mathbf{x}_k)$. Thus, S is a T_L-similarity relation. □

Let F be a fuzzy subset of X. The S-lower and S-upper approximations of F are denoted as:

$$\underline{S}(F)(\mathbf{x}_i) = \inf_{\mathbf{x}_j \in X} I_{T_L}\{S(\mathbf{x}_i, \mathbf{x}_j), F(\mathbf{x}_j)\}, \tag{8}$$

$$\overline{S}(F)(\mathbf{x}_i) = \sup_{\mathbf{x}_j \in X} T_L\{S(\mathbf{x}_i, \mathbf{x}_j), F(\mathbf{x}_j)\}, \tag{9}$$

where $I_{T_L}(a,b) = \min\{1, 1-a+b\}$ for every $a, b \in [0,1]$.

Like the approximation accuracy defined by Eq. (4), we have the following definition.

Definition 2: Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subset R^p$ be a given set of data points and F a fuzzy subset of X. The approximation accuracy of F is defined as:

$$\alpha_S(F) = \frac{|\underline{S}(F)|}{|\overline{S}(F)|}. \tag{10}$$

Here, $|A| = \Sigma_x A(x)$ is the cardinality of a fuzzy set A.

It is easy to prove that $\forall \mathbf{x}_i \in X$, $\underline{S}(F)(\mathbf{x}_i) \leq \overline{S}(F)(\mathbf{x}_i)$. So $\alpha_S(F) \leq 1$.

In general, S reflects the geometric structure of all data points in X. Intuitively, we can use the above concepts to estimate the proximity of two clusters.

Definition 3: Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subset R^p$ be a given set of data points and $(\mathbf{U}, V)$ a fuzzy partition. For two clusters F_i and F_j, the proximity of them is defined as:

$$P(F_i, F_j) = \frac{\frac{|F_i \cap F_j|}{|F_i| + |F_j|} \frac{1}{\alpha_S(F_i \cap F_j)}}{\| \mathbf{v}_i - \mathbf{v}_j \|} \tag{11}$$

$$= \frac{|F_i \cap F_j|}{(|F_i| + |F_j|)\alpha_S(F_i \cap F_j) \| \mathbf{v}_i - \mathbf{v}_j \|}, \tag{12}$$

where $F_i \cap F_j$ is defined as: $(F_i \cap F_j)(\mathbf{x}_i) = F_i(\mathbf{x}_i) \wedge F_j(\mathbf{x}_i)$.

In Eq. (11), the denominator indicates the separation strength of F_i and F_j. The numerator indicates the overlap strength of F_i and F_j, in which $\frac{1}{\alpha_S(F_i \cap F_j)}$ acts as a punishing factor. In general, a low value of $\alpha_S(F_i \cap F_j)$ suggests that there exists $X' \subseteq X$, $\forall \mathbf{x} \in X'$, there exists at least one data point close to $\mathbf{x}$ but its membership value belonging to $F_i \cap F_j$ is far different to that of $\mathbf{x}$. This means a low level of consistency of the overlapping part. So $\frac{1}{\alpha_S(F_i \cap F_j)}$ is introduced to assign more weight to less consistent overlapping part. A low value of $P(F_i, F_j)$ indicates low proximity of F_i and F_j.

Based on the above definition, we define the inter-cluster proximity of a fuzzy partition $(\mathbf{U}, V)$ as follows:

Definition 4: Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subset R^p$ be a given set of data points, $(\mathbf{U}, V)$ a fuzzy partition of X, and c the number of clusters. The proposed inter-cluster proximity index of $(\mathbf{U}, V)$ is defined as:

$$V_{proposed}(\mathbf{U}, V) = \frac{2}{c(c-1)} \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} P(F_i, F_j). \tag{13}$$

$V_{proposed}$ is the average inter-cluster proximity of all possible pairs of clusters. A low value of $V_{proposed}$ indicates low inter-cluster proximity of the corresponding fuzzy partition of X.

4. Experimental Results

In order to evaluate the performance of the proposed index, we applied $V_{proposed}$ and several well-known cluster validity indices, including the extended Xie-Beni index ($V_{XB,m}$) [2], [9], the Fukuyama-Sugeno index ($V_{FS,m}$) [3], the extended Kown index ($V_{Kwon,m}$) [10] and the Kim index (V_{Kim}) [5], to fuzzy partitions obtained from FCM for three data sets. Functional descriptions of the above indices are listed in Table 1.

The first two data sets are synthetic data sets, which are shown in Figs. 1 and 2 respectively. The synthetic data set 1 consists of 50 data points, with five well separated clusters. The synthetic data set 2 consists of 160 data points, with eight overlapping clusters.

The third one is a real data set: the IRIS data set [11]. It represents three categories of irises, with 50 samples per category. But it is known that two of the categories have substantial overlap while the third is well separated from the other two [9]. Thus, the most suitable cluster number is three or two.

For the mentioned data sets, we made several runs of FCM for different values of c. For a particular c and data set, FCM started from the same initial partition and ran for different values of m. We adopted Pal and Bezdek's suggestion [9]: $m \in [1.5, 2.5]$, $C_{min} = 2$ and $C_{max} \leq \sqrt{n}$. Thus, C_{max} was set to 7, 12 for the two synthetic data sets respectively and 12 for the IRIS data set. In all experiments, the distance function $\| \cdot \|$ was defined as Euclidean distance, and if an improvement in $J_m(\mathbf{U}, V)$ less than 10^{-5} was found, FCM stopped.

The experimental results are shown in Tables 2–4. One can see that:

(1) $V_{XB,m}$, V_{Kim} and $V_{proposed}$ correctly identify the optimal cluster number (denoted as c^* in the rest of this letter) of the synthetic data set 1 for all values of m, whereas $V_{FS,m}$ and $V_{Kwon,m}$ fail to do so.

(2) None of the five indices correctly identify c^* of the synthetic data set 2 for all values of m. V_{Kim} fails to find c^* for all values of m. In the other four indices, $V_{proposed}$ finds c^* on seven values of m, which shows the best performance. Although to some extent c=8 with m=1.5 can be viewed as an outlier under the assumption that the proposed index has monotonically decreasing feature, $V_{proposed}$ still finds c^* for the cases of $m \geq 2$. This result equals to the sub-optimal cluster number estimation obtained by the extended Kown index.

(3) All indices except $V_{FS,m}$ correctly identify c^* of the IRIS data sets for all values of m.

Overall, the proposed index can yield more desirable cluster number estimation in comparison to the other four indices.

Table 1 Four tested validity indices.

Index	Functional description
$V_{XB,m}$	$\frac{J_m(\mathbf{U},V)}{n \min_{i\neq j}\|v_i - v_j\|^2}$
$V_{FS,m}$	$J_m(\mathbf{U},V) - \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \parallel v_i - \bar{v}_1 \parallel^2$
$V_{Kwon,m}$	$\frac{J_m(\mathbf{U},V)+\frac{1}{c}\sum_{i=1}^{c}\|v_i-\bar{v}_2\|^2}{\min_{i\neq j}\|v_i-v_j\|^2}$
V_{Kim}	$\frac{2}{c(c-1)}\sum_{p\neq q}^{c}\sum_{j=1}^{n}(c(F_p(x_j) \wedge F_q(x_j))h(x_j))$

$\bar{v}_1 = \frac{1}{c}\sum_{i=1}^{c} v_i$, $\bar{v}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i$, $h(x_j) = -\sum_{i=1}^{c} F_i(x_j)\log F_i(x_j)$

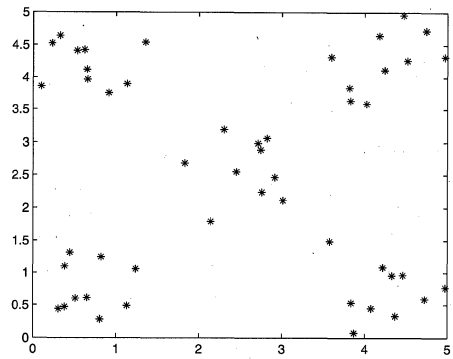

Fig. 1 Synthetic data set 1 (optimal cluster number is 5).

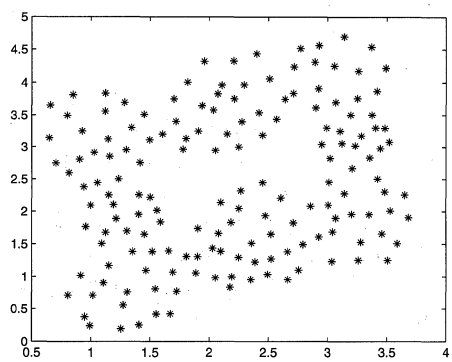

Fig. 2 Synthetic data set 2 (optimal cluster number is 8).

Table 2 Preferable values of c chosen by each index for Synthetic data set 1: c = 2–7.

m	$V_{XB,m}$	$V_{FS,m}$	$V_{Kwon,m}$	V_{Kim}	$V_{proposed}$
1.5	5	6	5	5	5
1.6	5	7	5	5	5
1.7	5	7	5	5	5
1.8	5	5	5	5	5
1.9	5	5	5	5	5
2.0	5	5	5	5	5
2.1	5	5	5	5	5
2.2	5	5	5	5	5
2.3	5	5	4	5	5
2.4	5	5	4	5	5
2.5	5	4	4	5	5

5. Conclusions

In this letter we propose a new index to evaluate the inter-cluster proximity of a fuzzy partition obtained from FCM. The proposed index is defined as the average proximity of all

Table 3 Preferable values of c chosen by each index for Synthetic data set 2: c = 2–12.

m	$V_{XB,m}$	$V_{FS,m}$	$V_{Kwon,m}$	V_{Kim}	$V_{proposed}$
1.5	4	12	4	12	8
1.6	4	12	4	12	12
1.7	4	12	4	12	12
1.8	4	12	4	2	12
1.9	4	9	4	2	12
2.0	8	8	8	2	8
2.1	8	9	8	2	8
2.2	8	8	8	2	8
2.3	8	8	8	2	8
2.4	8	8	8	2	8
2.5	12	8	8	2	8

Table 4 Preferable values of c chosen by each index for IRIS data set: c = 2–12.

m	$V_{XB,m}$	$V_{FS,m}$	$V_{Kwon,m}$	V_{Kim}	$V_{proposed}$
1.5	2	9	2	2	2
1.6	2	4	2	2	2
1.7	2	5	2	2	2
1.8	2	5	2	2	2
1.9	2	5	2	2	2
2.0	2	5	2	2	2
2.1	2	5	2	2	2
2.2	2	5	2	2	2
2.3	2	5	2	2	2
2.4	2	5	2	2	2
2.5	2	5	2	2	2

possible pairs of clusters. When quantifying the proximity of each pair of clusters, two kinds of information, the overlap and the separation of the two clusters, are considered. The former is quantified by using concepts of Fuzzy Rough sets theory and the latter by computing the distance between cluster centroids. Experimental results show that contrasted with some existing cluster validity indices involving only the membership value or the cluster centroids distance (including its variations) to measure the inter-cluster proximity, the proposed index provides a superior cluster number estimation. In future works, we plan to apply the basic ideas described in this letter to the cluster validity analysis for crisp clustering algorithms.

References

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

[2] X.L. Xie and G. Beni, "A validity measure for fuzzy clustering," IEEE Trans. Pattern Anal. Mach. Intell., vol.13, no.8, pp.841–847, Aug. 1991.

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

[4] D. Kim, K.H. Lee, and D. Lee, "On cluster validity index for estimation of the optimal number of fuzzy clusters," Pattern Recognit., vol.37, no.10, pp.2561–2574, Oct. 2004.

[5] Y. Kim, D. Kim, D. Lee, and K.H. Lee, "A cluster validation index for GK cluster analysis based on relative degree of sharing," Inf. Sci., vol.168, no.1-4, pp.225–242, Dec. 2004.

[6] Z. Pawlak, Rough sets-Theoretical aspects of reasoning about data, Kluwer Academic Publishers, Dordrecht, 1991.

[7] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," Int. J. Gen. Syst., vol.17, no.2-3, pp.191–209, March 1990.

[8] N.N. Morsi and M.M. Yakout, "Axiomatics for fuzzy rough set," Fuzzy Sets Syst. vol.100, no.1-3, pp.327–342, Nov. 1998.

[9] N.R. 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, Aug. 1995.

[10] S.H. Kwon, "Cluster validity index for fuzzy clustering," Electron. Lett., vol.34, no.22, pp.2176–2177, Sept. 1998.

[11] R.A. Fisher, "The use of multiple measurements in taxonomic problems," Ann. Eugenics, vol.7, no.2, pp.179–188, 1936.