IEICE TRANS. INF. & SYST., VOL.E88–D, NO.3 MARCH 2005

PAPER

Assessing the Quality of Fuzzy Partitions Using Relative Intersection

Dae-Won KIM[†a)], **Young-il KIM**[††], **Doheon LEE**[†], *and* **Kwang Hyung LEE**[†,††], ***Nonmembers***

SUMMARY In this paper, conventional validity indexes are reviewed and the shortcomings of the fuzzy cluster validation index based on inter-cluster proximity are examined. Based on these considerations, a new cluster validity index is proposed for fuzzy partitions obtained from the fuzzy *c*-means algorithm. The proposed validity index is defined as the average value of the relative intersections of all possible pairs of fuzzy clusters in the system. It computes the overlap between two fuzzy clusters by considering the intersection of each data point in the overlap. The optimal number of clusters is obtained by minimizing the validity index with respect to *c*. Experiments in which the proposed validity index and several conventional validity indexes were applied to well known data sets highlight the superior qualities of the proposed index.

key words: *cluster validity, fuzzy clustering, fuzzy c-means*

1. Introduction

Fuzzy clustering algorithms partition a data set into c homogeneous fuzzy clusters. Of the fuzzy clustering methods developed to date, the fuzzy c-means (FCM) algorithm [1] is the most widely used. The FCM method requires the number of clusters as an input, and the analysis result can vary greatly depending on the value chosen for this variable. However, in many cases the exact number of clusters in a data set is not known. In such cases, we can use a range of c values and then devise a validation index to determine the optimal number of clusters.

For this evaluation process, referred to as cluster validity, numerous validity indexes have been developed [1]–[7], [15]. Most of these indexes measure intra-cluster compactness and inter-cluster separation using cluster centroids. However, interpretation of inter-cluster separation of these indexes is problematic because such indexes quantify cluster separation based on the distance between cluster centroids only [15].

Recently, Kim et al. [15] proposed a fuzzy cluster validation index (v_P) based on inter-cluster proximity. Their index focuses on the degree to which pairs of clusters are separated by measuring the degree of overlap between clusters. Compared to conventional validity measures, the index of Kim et al. [15] showed superior performance when applied to a variety of well known data sets. However, it still suffers from the monotonic decreasing tendency with an increasing number of clusters. Moreover, it is sensitive to the choice of model parameters, and therefore determination of appropriate values for those parameters is crucial to the reliability of the index.

In this paper, the problems associated with conventional validity indexes are reviewed and their shortcomings are studied. Taking the problems of existing algorithms into account, a new cluster validity index for FCM is proposed that quantifies the relationship between each pair of clusters by calculating the relative intersection of two fuzzy sets. In this method, the intersection between two fuzzy clusters is computed by considering the degree of sharing of each datum in the overlap. Finally, the performance of the new validity measure is tested by applying it to well known data sets and comparing the results with those obtained using conventional validity indexes.

The remainder of this paper is organized as follows: Section 2 provides background information of fuzzy clustering and discusses previous work in cluster validity; Section 3 describes the formulation of the proposed validity index; Section 4 gives the results of experiments on a variety of data sets; and Section 5 presents our concluding remarks.

2. Fuzzy Cluster Validity Index

2.1 Fuzzy c-Means Algorithm

Fuzzy clustering algorithms generate a fuzzy partition given as a fuzzy partition matrix $U = [\mu_{ij}]$, where $\mu_{ij} = \mu_{\tilde{F}_i}(x_j)$ is the membership value of the data x_j belonging to the fuzzy cluster $\tilde{F}_i$. Fuzzy clustering algorithms are less prone to falling into local minima than crisp clustering algorithms because they make soft decisions at each iteration through the use of membership functions [8]–[12].

The most widely used fuzzy clustering algorithm is the FCM algorithm proposed by Bezdek [1]. This algorithm classifies a collection of data X into c homogeneous groups. The objective of FCM is to obtain a fuzzy c-partition $\tilde{F} = \{\tilde{F}_1, \ldots, \tilde{F}_c\}$ for the given number of clusters c and the given data $X = \{x_1, \ldots, x_n\}$ by minimizing the evaluation function J_m,

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

where $V = (v_1, \ldots, v_c)$ is a vector of the centroids of the

Manuscript received July 9, 2004.
Manuscript revised October 29, 2004.

[†]The authors are with the Department of BioSystems and Advanced Information Technology Research Center, KAIST, Guseong-dong, Yuseong-gu, Daejeon, Korea.
[††]The authors are with the Department of Electrical Engineering & Computer Science, KAIST, Guseong-dong, Yuseong-gu, Daejeon, Korea.
a) E-mail: dwkim@bisl.kaist.ac.kr

DOI: 10.1093/ietisy/e88–d.3.594

fuzzy clusters $(\tilde{F}_1, \ldots, \tilde{F}_c)$, $\|\cdot\|$ is a Euclidean norm, and m controls the fuzziness of membership of each datum. A fuzzy partition can be represented by (U, V). FCM tries to minimize $J_m(U, V : X)$ iteratively until no further improvement is possible.

2.2 Conventional Cluster Validity Indexes

Cluster validity indexes are used to establish which partition best explains the unknown cluster structure in a given data set [16]. FCM is run over a range of c values, $2, \ldots, c_{max}$, and the resulting fuzzy partition is evaluated with the validity indexes to identify the optimal number of clusters. Usually, $c_{max} \approx \sqrt{n}$ is used [4].

Bezdek proposed two cluster validity indexes for fuzzy clustering [2], [3]. These indexes, which are referred to as the Partition Coefficient (v_{PC}) and Partition Entropy (v_{PE}), are defined as

$$v_{PC} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^2}{n}, \tag{2}$$

$$v_{PE} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}[\mu_{ij}\log_a(\mu_{ij})]. \tag{3}$$

The optimal fuzzy partition is obtained by maximizing v_{PC} (or minimizing v_{PE}) with respect to $c = 2, \ldots, c_{max}$.

Xie and Beni proposed a validity index (v_{XB}) that focuses on two properties: compactness and separation [5]. v_{XB} is defined as

$$v_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\|x_j - v_i\|^2}{n(\underbrace{min}_{i\neq k}\|v_i - v_k\|^2)} \tag{4}$$

In this equation, the numerator is the sum of the compactness of each fuzzy cluster and the denominator is the minimal separation between fuzzy clusters. The optimal fuzzy partition is obtained by minimizing V_{XB} with respect to $c = 2, \ldots, c_{max}$.

v_{XB} decreases monotonically as $c \to n$. Kwon extended v_{XB} to eliminate this decreasing trend [6] by adding a penalty value to the numerator of v_{XB}. Kwon's index (v_K) is given as

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

Rezaee combined a measure of the average scatter of c clusters, $Scat(c)$, with the distance functional, $Dist(c)$ [7]. Rezaee's validity index (v_{CWB}) is defined as

$$\begin{aligned} v_{CWB} &= \alpha Scat(c) + Dist(c) \\ &= \alpha\frac{\sum_{i=1}^{c}\|\sigma(v_i)\|}{c\|\sigma(X)\|} \\ &\quad + \frac{D_{max}}{D_{min}}\sum_{k=1}^{c}\left(\sum_{z=1}^{c}\|v_k - v_z\|\right)^{-1} \end{aligned} \tag{6}$$

where $\sigma(v_i)$ is the fuzzy variance of the i-th cluster, and $\sigma(X)$ represents the variance of the data set X. D_{max} and D_{min} are the maximum and minimum distances between the cluster centroids respectively.

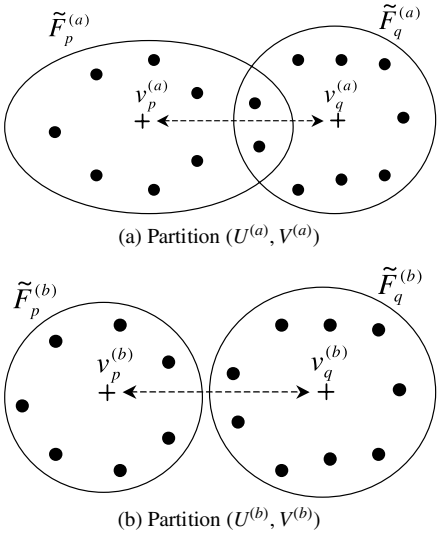
As recently pointed out by Kim et al. [15], most validity indexes focus only on the compactness and the variation of the intra-cluster distance [5]–[7]. Some indexes, for example v_{XB}, v_K and v_{CWB}, use the strength of separation between clusters; however, interpretation of these indexes is problematic because they quantify cluster separation based only on the distance between cluster centroids [15]. The problem addressed by Kim et al. is demonstrated in Fig. 1, which shows two different fuzzy partitions $(U^{(a)}, V^{(a)})$ and $(U^{(b)}, V^{(b)})$ with the same distance between cluster centroids for some data. In this figure, even though $(U^{(b)}, V^{(b)})$ provides a better partitioning than $(U^{(a)}, V^{(a)})$, conventional validity indexes cannot discriminate between these two fuzzy partitions because they only use distance between the cluster centroids.

To tackle this problem, Kim et al. proposed a new validity index that exploits the geometric properties of fuzzy clusters [15]. Their approach is based on an inter-cluster proximity index (v_P) between fuzzy sets. v_P is defined as follows,

$$v_P = \frac{2}{c(c-1)}\sum_{p\neq q}^{c}\left[\sum_{\mu}\sum_{j=1}^{n}\delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q)\omega(x_j)\right] \tag{7}$$

where $\delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q) = 1$ if $\mu_{\tilde{F}_p}(x_j) \wedge \mu_{\tilde{F}_q}(x_j) \geq \mu$ and 0 otherwise.

$\delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q)$ determines whether two clusters are proximate at the membership degree μ for data x_j. It returns a proximity of 1.0 when the membership degrees of both clusters are greater than μ, and returns 0.0 otherwise. $\omega(x_j) \in [0.0, 1.0]$ is a weight function that is selected to give more weight to vague data and less weight to clearly classi-



(a) Partition $(U^{(a)}, V^{(a)})$

(b) Partition $(U^{(b)}, V^{(b)})$

Fig. 1 Two different fuzzy partitions $(U^{(a)}, V^{(a)})$ and $(U^{(b)}, V^{(b)})$ with the same distance between cluster centroids for some data.

fied data. In the experiments of Kim et al. [15], $\omega(x_j)$ was assigned a value of 0.1 ($\mu_{\tilde{F}_i}(x_j) \geq 0.8$), 0.4 ($0.7 \leq \mu_{\tilde{F}_i}(x_j) < 0.8$), and 0.7 ($0.6 \leq \mu_{\tilde{F}_i}(x_j) < 0.7$) for any $\tilde{F}_i \in \tilde{F}$; otherwise, $\omega(x_j)$ was assigned a value of 1.0.

3. The Proposed Validity Index Using Relative Intersection

3.1 Motivation

The validity index v_P uses the proximity between fuzzy clusters, and determines this proximity based on the similarity between fuzzy clusters. It focuses on the extent to which the clusters in each cluster pair are separated by measuring the degree of overlap between clusters. Furthermore, by using $\omega(x_j)$ for each datum, v_P can concentrate more on highly overlapped data in the computation of the validity index. The optimal number of clusters is obtained by minimizing v_P with respect to $c = 2, \ldots, c_{max}$. When applied to well known test data sets, v_P showed an excellent ability to find the optimal number of clusters and to be more reliable than other indexes [15].

Although v_P has shown its superior validation performance to other indexes, it has two shortcomings: (1) a monotonic decreasing tendency for larger values of c, and (2) the choice of weight function, $\omega(x_j)$.

Let us consider the first issue. Like the problems of the conventional validity indexes pointed out by Pal [4] and Kwon [6], the proximity index v_P still suffers from the monotonic decreasing tendency when c approaches to larger values. This is because v_P is sensitive to the choice of μ. As seen in Eq. (7), the proximity of two fuzzy clusters is calculated in each μ level where $\{\mu\}$ is not explicitly formulated [15].

When $\{\mu\}$ is an infinite set like $\mu \in (0.0, 1.0]$, $v_P = \infty$ for all c values and the fuzzy cluster validation does not give a meaningful result. When $\{\mu\}$ is an finite set like $\{\mu\} = \{1/d, 2/d, \ldots, d/d\}$ where d is a discretization unit, the proximity of two fuzzy clusters can be rewritten from Eq. (7),

$$
\begin{aligned}
&\sum_{\mu\in\{1/d,\ldots,d/d\}} \sum_{j=1}^{n} \delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q)\omega(x_j) \\
&= \sum_{j=1}^{n} \delta(x_j, 1/d : \tilde{F}_p, \tilde{F}_q)\omega(x_j) \\
&\quad + \sum_{j=1}^{n} \delta(x_j, 2/d : \tilde{F}_p, \tilde{F}_q)\omega(x_j) \\
&\quad + \ldots + \sum_{j=1}^{n} \delta(x_j, d/d : \tilde{F}_p, \tilde{F}_q)\omega(x_j) \\
&= \sum_{\mu\in\{1/d,\ldots,d/d\}} \delta(x_1, \mu : \tilde{F}_p, \tilde{F}_q)\omega(x_1) \\
&\quad + \sum_{\mu\in\{1/d,\ldots,d/d\}} \delta(x_2, \mu : \tilde{F}_p, \tilde{F}_q)\omega(x_2) \\
&\quad + \ldots + \sum_{\mu\in\{1/d,\ldots,d/d\}} \delta(x_n, \mu : \tilde{F}_p, \tilde{F}_q)\omega(x_n) \\
&= \left\lfloor \frac{\mu_{\tilde{F}_p}(x_1) \wedge \mu_{\tilde{F}_q}(x_1)}{1/d} \right\rfloor \omega(x_1) \\
&\quad + \left\lfloor \frac{\mu_{\tilde{F}_p}(x_2) \wedge \mu_{\tilde{F}_q}(x_2)}{1/d} \right\rfloor \omega(x_2) \\
&\quad + \ldots + \left\lfloor \frac{\mu_{\tilde{F}_p}(x_n) \wedge \mu_{\tilde{F}_q}(x_n)}{1/d} \right\rfloor \omega(x_n) \\
&= \sum_{j=1}^{n} \left\lfloor d \cdot \left[\mu_{\tilde{F}_p}(x_j) \wedge \mu_{\tilde{F}_q}(x_j) \right] \right\rfloor \omega(x_j)
\end{aligned}
$$

where the floor function $\lfloor \tau \rfloor$ gives the largest integer less than or equal to τ.

Let us define $\mu_{max1}(x_j)$, $\mu_{max2}(x_j)$ and $\mu_{rest}(x_j)$ as follows:

$$
\begin{aligned}
&\mu_{max1}(x_j) = \max(\{\mu_{ij} | 1 \leq i \leq c\}), \\
&\mu_{max2}(x_j) = \max(\{\mu_{ij} | 1 \leq i \leq c\} - \{\mu_{max1}(x_j)\}), \\
&\mu_{rest}(x_j) = 1 - \mu_{max1}(x_j) - \mu_{max2}(x_j).
\end{aligned}
$$

From Fig. 2, it is evident that

$$
0 \leq \max_{p\neq q} \left[\mu_{\tilde{F}_p}(x_j) \wedge \mu_{\tilde{F}_q}(x_j) \right] \leq \frac{1 - \mu_{rest}(x_j)}{2} \leq \frac{1}{2} \tag{8}
$$

Hence, v_P can be rewritten as

$$
\begin{aligned}
v_P &= \frac{2}{c(c-1)} \sum_{p\neq q}^{c} \sum_{j=1}^{n} \left\lfloor d \cdot \left[\mu_{\tilde{F}_p}(x_j) \wedge \mu_{\tilde{F}_q}(x_j) \right] \right\rfloor \omega(x_j) \\
&\leq \frac{2}{c(c-1)} \sum_{p\neq q}^{c} \sum_{j=1}^{n} \left\lfloor d \cdot \frac{1 - \mu_{rest}(x_j)}{2} \right\rfloor \omega(x_j)
\end{aligned} \tag{9}
$$

Because the FCM algorithm is a probabilistic clustering algorithm [16], the membership values are assigned relative to each other. This means that as c increases, $\mu_{rest}(x_j)$ tends to increase. Hence we see from Eq. (9) that, for a finite level set $\{\mu\}$, v_P tends to decrease as the value of c increases. This tendency becomes very marked for level sets with higher resolutions (i.e., larger values of d), as will be

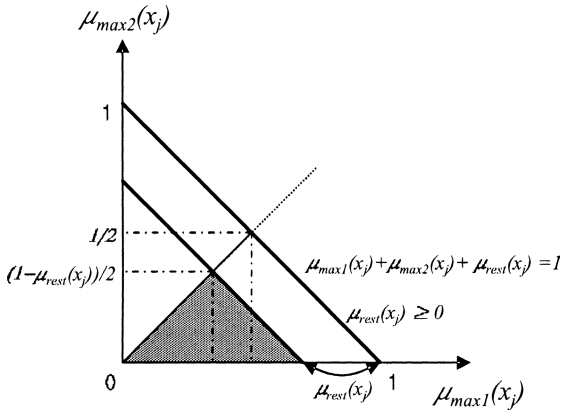


Fig. 2 The shaded region shows $\mu_{max1}(x_j) + \mu_{max2}(x_j) + \mu_{rest}(x_j) = 1$, $\mu_{max1}(x_j) \geq \mu_{max2}(x_j)$ and $\mu_{rest}(x_j) \geq 0$.

seen in Sect. 4 (see Tables 2, 4 and 7).

Another shortcoming of v_P lies on the choice of the weight function, $\omega(x_j)$. Kim et al. [15] used the following simple step function in their experiments:

$$\omega(x_j) = \begin{cases} 0.1 & \text{if } \mu_{\tilde{F}_q}(x_j) \geq 0.8 \\ 0.4 & \text{if } 0.7 \leq \mu_{\tilde{F}_q}(x_j) \leq 0.8 \\ 0.7 & \text{if } 0.6 \leq \mu_{\tilde{F}_q}(x_j) \leq 0.7 \\ 1.0 & \text{otherwise} \end{cases} \tag{10}$$

However, this type of fixed $\omega(x_j)$ does not reflect the dependency of $\mu_{\tilde{F}_i}(x_j)$ on different c values discussed above. Thus a systematic way of weighting technique for fuzzy partitions obtained from different c values is required.

3.2 Relative Intersection of Two Fuzzy Clusters

To solve the addressed problems, in the present study we exploit two notions: a relative intersection and an entropy. Firstly, to eliminate the monotonic decreasing tendency for increasing c values, we employ the notion of relative intersection that is defined as the weighted sum of the relative intersection for all data. Secondly, to calculate the weights of data clustered for different c values, we use an entropy function instead of the simple step function. Then, we propose a new validity index (v_{RI}) that is defined as the average value of the relative intersections of all possible pairs of fuzzy clusters in the system.

A relative intersection of two fuzzy clusters at each datum x_j is calculated before computing the total relative intersection of two fuzzy clusters. Let $\tilde{F}_p$ and $\tilde{F}_q$ be two fuzzy clusters belonging to a fuzzy partition (U, V) and c be the number of clusters. Then the relative intersection of two fuzzy clusters $\tilde{F}_p$ and $\tilde{F}_q$ at x_j is defined as

$$I_R(x_j : \tilde{F}_p, \tilde{F}_q) = \frac{\mu_{\tilde{F}_p}(x_j) \wedge \mu_{\tilde{F}_q}(x_j)}{(1/c)\sum_{i=1}^{c} \mu_{\tilde{F}_i}(x_j)} \tag{11}$$

In Eq. (11), the numerator is the intersection of $\tilde{F}_p$ and $\tilde{F}_q$ at x_j, indicating $\min(\mu_{\tilde{F}_p}(x_j), \mu_{\tilde{F}_q}(x_j))$, and the denominator is the average membership value of x_j over c fuzzy clusters. Use of the relative intersection in validating fuzzy partitions makes it possible to observe the relative quality of two fuzzy clusters at the viewpoint of the whole partition.

Since $\sum_{i=1}^{c} \mu_{\tilde{F}_i}(x_j) = 1$ in FCM, $I_R(x_j : \tilde{F}_p, \tilde{F}_q)$ can be rewritten as

$$I_R(x_j : \tilde{F}_p, \tilde{F}_q) = c \cdot \left[\mu_{\tilde{F}_p}(x_j) \wedge \mu_{\tilde{F}_q}(x_j)\right] \tag{12}$$

From Eqs. (11) and (12), we can see that although $\mu_{\tilde{F}_p}(x_j) \wedge \mu_{\tilde{F}_q}(x_j)$ tends to decrease (i.e., $\mu_{rest}(x_j)$ increases) for larger number of clusters, the value of I_R is compensated by the number of clusters c. Thus, the relative intersection I_R can avoid the monotonic decreasing tendency as the number of clusters increases.

Figure 3 depicts two intersection values at datum x_j between two fuzzy clusters. Two fuzzy partitions, $U^{(a)}$ and $U^{(b)}$, are shown in the figure. For simplicity, we assume that

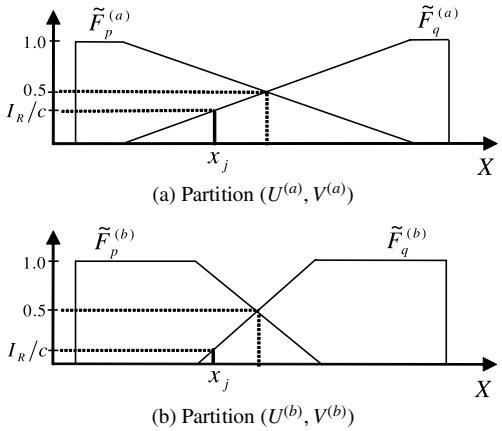


Fig. 3 Intersection of two fuzzy clusters: (a) partition $(U^{(a)}, V^{(a)})$, (b) partition $(U^{(b)}, V^{(b)})$.

each fuzzy cluster is represented by a triangular fuzzy set. Given the membership degrees $\mu_{\tilde{F}_p^{(a)}}(x_j)$ and $\mu_{\tilde{F}_q^{(a)}}(x_j)$, relative intersection $I_R(x_j : \tilde{F}_p^{(a)}, \tilde{F}_q^{(a)})$ is obtained by the product of c and $\mu_{\tilde{F}_p^{(a)}}(x_j) \wedge \mu_{\tilde{F}_q^{(a)}}(x_j)$. When we compare two relative intersection values $I_R(x_j : \tilde{F}_p^{(a)}, \tilde{F}_q^{(a)})$ and $I_R(x_j : \tilde{F}_p^{(b)}, \tilde{F}_q^{(b)})$ for the partitions $U^{(a)}$ and $U^{(b)}$, $I_R(x_j : \tilde{F}_p^{(b)}, \tilde{F}_q^{(b)})$ is given a lower value than $I_R(x_j : \tilde{F}_p^{(a)}, \tilde{F}_q^{(a)})$. This indicates that the partition $U^{(b)}$ more clearly classified the datum x_j than $U^{(a)}$. By considering the intersections for all $x_j \in X$, we can see that the partition $U^{(a)}$ contains a greater number of vague data than $U^{(b)}$, and hence $U^{(b)}$ is the better of the two partitions.

When assessing fuzzy clusters, it is of importance to use information regarding how vaguely (unclearly) the datum x_j is classified over c different clusters. The more clearly classified data the fuzzy clusters have, the better the quality of the fuzzy partition. To quantify how clearly the datum x_j is classified, the entropy of x_j is exploited in the present study, which is defined as

$$E(x_j) = -\sum_{i=1}^{c} \mu_{\tilde{F}_i}(x_j) \log_e \mu_{\tilde{F}_i}(x_j) \tag{13}$$

Here, $E(x_j)$ is the entropy of datum x_j and $\mu_{\tilde{F}_i}(x_j)$ is the membership value with which x_j belongs to cluster $\tilde{F}_i$. By the properties of entropy, the greater the vagueness of the classified datum x_j, the greater the entropy of x_j. By considering this entropy, vague data are given more weight than clearly classified data when assessing fuzzy clusters. Unlike $\omega(x_j)$ used in v_P [15], $E(x_j)$ reflects the dependency of $\mu_{\tilde{F}_i}(x_j)$ with respect to different c values. This approach makes it possible to focus more on the highly-overlapped data in the computation of the validity index than other indexes do.

Definition 1: Let $\tilde{F}_p$ and $\tilde{F}_q$ be two fuzzy clusters belonging to a pattern matrix U. Let $I_R(x_j : \tilde{F}_p, \tilde{F}_q)$ be a relative intersection at datum x_j between $\tilde{F}_p$ and $\tilde{F}_q$. And let $E(x_j)$ be an entropy of x_j in fuzzy clusters. Then, the relative intersection of fuzzy clusters $\tilde{F}_p$ and $\tilde{F}_q$ is defined as

$$I_R(\tilde{F}_p, \tilde{F}_q) = \sum_{j=1}^{n} I_R(x_j : \tilde{F}_p, \tilde{F}_q) E(x_j), \tag{14}$$

$I_R(\tilde{F}_p, \tilde{F}_q)$ is formulated as the weighted summation of $I_R(x_j : \tilde{F}_p, \tilde{F}_q)$ for all data in X. $E(x_j)$ means the degree of vagueness for each datum x_j. A small value of $I_R(\tilde{F}_p, \tilde{F}_q)$ indicates that $\tilde{F}_p$ has little intersection with $\tilde{F}_q$, and therefore, two fuzzy clusters $\tilde{F}_p$ and $\tilde{F}_q$ are well-classified.

3.3 The Proposed Validity Index

Now we propose a new validity index based on the relative intersection. The proposed validity index is the average value of the relative intersections of all possible pairs of fuzzy clusters in the system.

Definition 2: Let $\tilde{F}_p$ and $\tilde{F}_q$ be two fuzzy clusters belonging to a fuzzy partition (U, V) and c be the number of clusters. Let $I_R(\tilde{F}_p, \tilde{F}_q)$ be the relative intersection of two fuzzy clusters. Then the proposed validity index v_{RI} is defined as

$$v_{RI}(c, U) = \frac{2}{c(c-1)} \sum_{p \neq q}^{c} I_R(\tilde{F}_p, \tilde{F}_q) \tag{15}$$

Thus, v_{RI} is defined as the average value of the relative intersections of $\frac{c(c-1)}{2}$ pairs of clusters, where the relative intersection of each cluster pair is defined as the weighted sum of the relative intersection at x_j of two clusters in the pair. Hence, the less overlap there is in a fuzzy partition, and the less vague the data points in that overlap, the lower the value of $v_{RI}(c, U)$. The optimal number of clusters is obtained by minimizing $v_{RI}(c, U)$ over the range of c values, $2, \ldots, c_{max}$. The procedure for finding the optimal number of clusters (or the optimal fuzzy partition) obtained through the FCM algorithm using v_{RI} is described above.

Algorithm 1 Find the optimal number of clusters

Input: data $X = \{x_1, \ldots, x_n\}$, maximum clusters c_{max}
termination criterion ϵ, fuzziness m
Output: the optimal number of clusters c_{opt}
Procedure:

```
c_max ← √n, m ← 2.0, ε ← 0.001, J(0) ← ∞;
for (c ← 2; c < c_max; c++) do
    t ← 1;
    Initialize the cluster centroids V(t);
    Compute the pattern matrix U(t);
    while |J(t) − J(t − 1)| > ε do
        t ← t + 1;
        Update the cluster centroids V(t);
        Update the pattern matrix U(t);
    end while
    Compute the validity index v_RI(c, U);
end for
Find the optimal c: c_opt ← arg_c min v_RI(c, U);
```

4. Experiments

To test the performance of v_{RI}, we used it to determine the optimal cluster numbers in seven well known data sets and compared the results with those obtained using v_{PC}, v_{PE}, v_{XB}, v_K, and v_P with three different level sets ($v_P^{0.1}$ with $\mu \in \{0.1, 0.2, \ldots, 1.0\}$, $v_P^{0.01}$ with $\mu \in \{0.01, 0.02, \ldots, 1.0\}$ and $v_P^{0.001}$ with $\mu \in \{0.001, 0.002, \ldots, 1.0\}$).

The data sets used for these experiments were X30 [14], Bensaid [13], AD-2 (2-clusters), AD-9 (9-clusters), AD-3D (4 clusters in 3D), a superset of Starfield [5], [15], and Iris [4]. The raw (unscaled) data were used without normalization. The parameters of the FCM were set as follows: termination criterion $\epsilon = 0.001$, weighting exponent $m = 2.0$, and Euclidean norm. Initial centroids were selected randomly. For the evaluation of validity indexes, $c_{max} \approx \sqrt{n}$ was used [4]. For $v_P^{0.1}$, $v_P^{0.01}$ and $v_P^{0.001}$, $\omega(x_j)$ in the proximity function was assigned a value of 0.1 ($\mu_{\tilde{F}_i}(x_j) \geq 0.8$), 0.4 ($0.7 \leq \mu_{\tilde{F}_i}(x_j) < 0.8$), and 0.7 ($0.6 \leq \mu_{\tilde{F}_i}(x_j) < 0.7$) for any $\tilde{F}_i \in \tilde{F}$; otherwise, $\omega(x_j)$ was set to 1.0, as per the work of Kim et al. Figs. 4–8 show scatter plots of five of the seven data sets used in the experiment, and Tables 1–7 show the results of the evaluation of each cluster validation index. Optimal c values are shown in bold face in the tables.

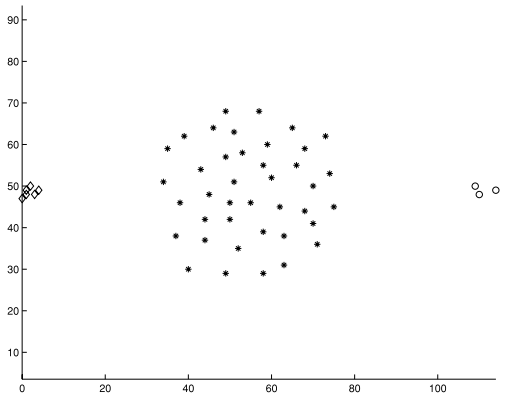


Fig. 4 Bensaid data set. 49 data points. Optimal cluster number is 3.

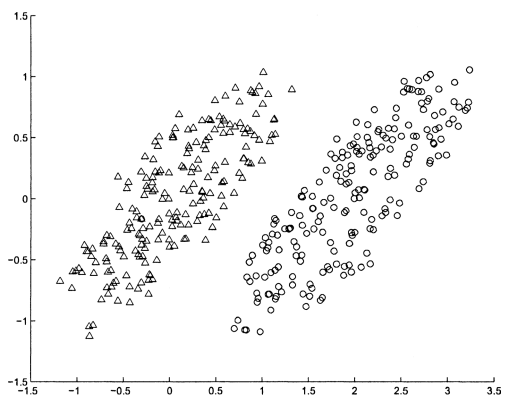


Fig. 5 AD-2 data set. 400 data points. Optimal cluster number is 2.

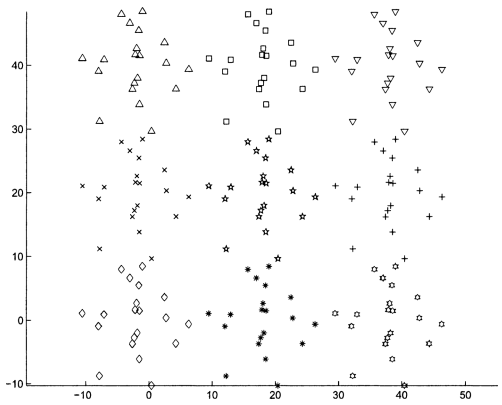


Fig. 6 AD-9 data set. 180 data points. Optimal cluster number is 9.

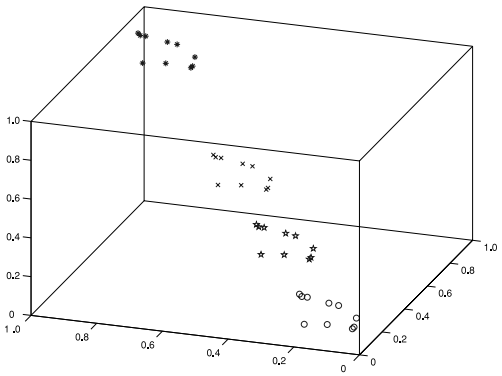


Fig. 7 AD-3D data set. 40 data points. Optimal cluster number is 4.

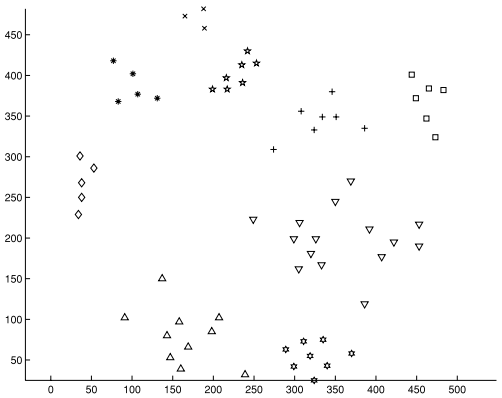


Fig. 8 Starfield data set. 66 data points. Optimal cluster number is 9.

The data points of the X30 [14] are arranged in three compact and well-separated clusters, each containing 10 points. For the X30, all of the validity indexes except v_K correctly identified the optimal number of clusters (Table 1).

For the Bensaid data set [13], which consists of 49 data points that are distributed in three compact and well-separated clusters containing different numbers of points (see scatter plot in Fig. 4), five cluster validity indexes including the proposed index v_{RI} correctly recognized the optimal $c = 3$. $v_P^{0.01}$ and $v_P^{0.001}$ show the tendency to decrease with increasing c (Table 2). As pointed out in the previous section, v_P is sensitive to the choice of the level set $\{\mu\}$.

The third data set is the AD-2 data set, which consists of 400 data points that are distributed in two separated clusters. Figure 5 shows the scatter plot of this data, each cluster contains 200 data points. The validation results of the indexes for $c = 2, 3, \ldots, c_{max} = 17$ are listed in Table 3. For this data set, v_{PC}, v_{PE}, v_{XB}, v_K, and v_{RI} correctly identified the optimal number of clusters. In contrast, v_{CWB}, $v_P^{0.1}$, $v_P^{0.01}$, and $v_P^{0.001}$ failed to detect the optimal number of clusters. It is observed that v_P showed the tendency to decrease with increasing c (Table 3).

The fourth data set, the AD-9, was constructed by generating 20 data points with a normal distribution and then creating nine copies of this cluster of 20 points and placing these copies at adjacent positions. Figure 6 shows the scatter plot of this data, which is distributed into nine clusters. The validation results of each index for $c = 2, 3, \ldots, c_{max} = \sqrt{n} \approx 13$ are listed in Table 4. For this data set, v_{PC}, v_{PE}, $v_P^{0.01}$ and $v_P^{0.001}$ failed to detect the optimal number of clusters, and $v_P^{0.01}$ and $v_P^{0.001}$ showed the tendency to decrease with increasing c (Table 4).

Figure 7 shows the AD-3D data set in which ten data points were generated randomly in three-dimensional space and then copied to four positions in a diagonal arrangement. Three of the clusters were placed adjacent to each other, and one cluster was separated from them. Table 5 shows the results obtained using the various validity indexes with respect to $c = 2, 3, \ldots, c_{max} = \sqrt{n} \approx 6$. Only $v_P^{0.1}$, $v_P^{0.01}$, $v_P^{0.001}$ and v_{RI} successfully detected the optimal number of clusters. In contrast, v_{PC}, v_{PE}, v_{XB}, and v_K determined the optimal value to be $c = 2$, and v_{CWB} yielded optimal partitions at $c = 3$. v_{XB} and v_K could not detect the optimal number of clusters because they use only the distance between centroids to measure the separation: the maximum separation is obtained when the number of clusters is two.

Figure 8 shows the scatter plot of a superset of the Starfield data set [5], [15]. The data set has 66 data points

Table 1 Cluster validity values for the X30 data set. Optimal cluster number is 3.

c	v_{PC}	v_{PE}	v_{XB}	v_K	v_{CWB}	$v_P^{0.1}$	$v_P^{0.01}$	$v_P^{0.001}$	v_{RI}
c=2	0.91	0.08	0.04	**1.45**	5.8	7.20	34.00	313.20	1.82
c=3	**0.97**	**0.04**	**0.02**	1.82	**2.47**	**6.00**	**8.20**	**42.33**	**0.19**
c=4	0.93	0.07	0.08	11.97	2.60	9.67	24.17	190.10	1.06
c=5	0.87	0.12	2.15	171.76	7.57	9.98	30.94	252.24	2.09

Table 2 Cluster validity values for the Bensaid data set. Optimal cluster number is 3.

c	v_{PC}	v_{PE}	v_{XB}	v_K	v_{CWB}	$v_P^{0.1}$	$v_P^{0.01}$	$v_P^{0.001}$	v_{RI}
c=2	0.72	**0.19**	0.24	11.92	0.84	112.60	993.20	9793.40	21.44
c=3	**0.75**	0.20	**0.07**	**4.12**	0.62	**39.87**	281.67	2706.47	**13.06**
c=4	0.67	0.26	0.22	15.06	0.55	63.00	378.50	3585.63	17.399
c=5	0.66	0.30	0.12	8.70	0.46	58.92	302.28	2835.160	19.50
c=6	0.63	0.33	0.10	7.92	**0.43**	56.07	276.25	2532.97	21.55
c=7	0.61	0.36	0.10	9.20	0.44	58.74	**251.92**	**2273.96**	23.69

Table 3 Cluster validity values for the AD-2 data set. Optimal cluster number is 2.

c	v_{PC}	v_{PE}	v_{XB}	v_K	v_{CWB}	$v_P^{0.1}$	$v_P^{0.01}$	$v_P^{0.001}$	v_{RI}
c=2	**0.80**	**0.14**	**0.11**	**42.67**	6.24	727.60	6325.00	62396.00	**120.48**
c=3	0.69	0.24	0.13	52.98	5.59	606.00	4642.68	45040.60	160.78
c=4	0.67	0.28	**0.11**	44.46	**5.05**	**432.33**	2799.40	26584.13	149.56
c=5	0.62	0.34	0.16	67.65	5.65	458.74	2537.00	23746.40	175.62
c=6	0.56	0.40	0.32	134.03	7.32	528.43	2696.40	24981.21	227.93
c=7	0.57	0.40	0.22	90.67	6.73	462.38	1860.08	16784.63	185.85
c=8	0.55	0.43	0.15	63.23	6.14	487.54	1712.71	15090.92	191.39
c=9	0.53	0.45	0.14	60.83	6.33	498.28	1600.65	13931.18	204.65
c=10	0.51	0.48	0.17	75.97	7.05	515.95	1511.73	12946.96	218.96
c=11	0.51	0.50	0.16	68.62	7.04	516.93	1380.91	11548.57	216.67
c=12	0.50	0.51	0.26	117.59	8.69	520.07	1291.93	10606.14	222.97
c=13	0.47	0.56	0.39	171.22	10.85	541.48	1347.51	11051.56	264.47
c=14	0.46	0.57	0.20	89.76	8.45	552.46	1232.33	9758.08	250.34
c=15	0.46	0.58	0.39	175.05	11.80	545.69	1174.88	9284.43	262.74
c=16	0.47	0.57	0.21	95.88	9.62	531.60	1051.46	8047.13	241.95
c=17	0.47	0.58	0.18	85.46	9.33	534.65	**996.19**	**7402.19**	238.20

Table 4 Cluster validity values for the AD-9. Optimal cluster number is 9.

c	v_{PC}	v_{PE}	v_{XB}	v_K	v_{CWB}	$v_P^{0.1}$	$v_P^{0.01}$	$v_P^{0.001}$	v_{RI}
c=2	**0.68**	**0.21**	0.34	60.78	0.82	561.00	4930.40	48718.60	97.38
c=3	0.61	0.30	0.14	25.32	0.61	428.00	3275.33	31910.73	111.96
c=4	0.59	0.34	0.09	16.73	0.49	305.23	2344.90	22648.83	120.36
c=5	0.56	0.39	0.17	32.65	0.47	269.36	1811.28	17230.52	128.38
c=6	0.54	0.42	0.13	25.17	0.43	257.44	1501.31	14115.43	134.65
c=7	0.56	0.42	0.11	20.62	0.39	216.03	1086.17	9993.82	119.36
c=8	0.58	0.42	0.08	15.04	0.35	189.44	826.55	7438.34	108.53
c=9	0.58	0.43	**0.06**	**12.18**	**0.33**	**186.52**	650.75	5663.74	**90.98**
c=10	0.56	0.46	0.20	40.63	0.41	201.27	645.70	5543.20	99.73
c=11	0.54	0.48	0.27	55.28	0.45	208.14	613.69	5182.65	105.12
c=12	0.53	0.50	0.18	37.47	0.41	224.86	607.07	5094.46	110.18
c=13	0.51	0.53	0.17	36.70	0.41	226.89	**581.07**	**4793.49**	115.89

Table 5 Cluster validity values for the AD-3D data set. Optimal cluster number is 4.

c	v_{PC}	v_{PE}	v_{XB}	v_K	v_{CWB}	$v_P^{0.1}$	$v_P^{0.01}$	$v_P^{0.001}$	v_{RI}
c=2	**0.83**	**0.12**	**0.07**	**2.87**	8.16	77.60	682.60	6722.60	11.27
c=3	0.80	0.16	0.07	3.66	**7.92**	30.93	198.47	1902.33	7.18
c=4	0.78	0.20	0.09	5.70	10.14	**12.37**	**49.27**	**446.83**	**5.62**
c=5	0.71	0.26	0.45	30.57	18.98	23.18	87.94	806.46	7.51
c=6	0.64	0.32	0.34	24.72	18.66	40.23	148.23	1327.60	11.57

Table 6 Cluster validity values for the Starfield data set. Optimal cluster number is 9.

c	v_{PC}	v_{PE}	v_{XB}	v_K	v_{CWB}	$v_P^{0.1}$	$v_P^{0.01}$	$v_P^{0.001}$	v_{RI}
c=2	**0.73**	**0.18**	0.24	16.04	0.38	148.60	1264.40	12468.20	26.89
c=3	0.66	0.26	0.12	**8.29**	0.27	93.33	721.53	6987.27	29.14
c=4	0.62	0.32	0.12	8.72	0.23	96.17	698.90	6748.80	35.90
c=5	0.63	0.34	0.16	11.68	0.20	72.72	450.82	4240.20	32.90
c=6	0.62	0.35	0.17	13.42	0.18	73.31	405.64	3760.07	34.22
c=7	0.66	0.33	0.11	9.61	0.14	57.83	269.20	2451.04	28.66
c=8	0.67	0.33	0.09	8.33	0.13	54.01	220.44	1959.08	27.05
c=9	0.70	0.31	**0.08**	8.40	**0.12**	**42.25**	**138.33**	**1192.70**	**20.82**
c=10	0.63	0.40	0.45	40.30	0.15	53.37	174.01	1502.78	29.82

Table 7 Cluster validity values for Iris data set. Optimal cluster number is 2.

c	v_{PC}	v_{PE}	v_{XB}	v_K	v_{CWB}	$v_P^{0.1}$	$v_P^{0.01}$	$v_P^{0.001}$	v_{RI}
c=2	**0.89**	**0.09**	**0.05**	**8.39**	5.04	**99.60**	741.80	7203.20	**17.43**
c=3	0.78	0.17	0.14	21.99	**4.46**	135.73	896.20	8580.20	31.84
c=4	0.71	0.24	0.20	32.43	4.77	150.10	868.63	8180.87	43.97
c=5	0.63	0.31	0.40	68.14	5.35	174.60	800.14	7338.76	48.90
c=6	0.57	0.36	0.74	128.81	6.55	189.33	752.20	6771.81	57.11
c=7	0.53	0.41	1.54	268.73	8.95	204.69	748.12	6605.32	67.24
c=8	0.52	0.44	0.44	84.43	6.33	203.25	649.00	5625.11	67.83
c=9	0.49	0.47	0.77	151.23	7.93	212.80	618.19	5176.12	72.60
c=10	0.47	0.50	1.27	252.76	9.97	215.95	583.43	4800.84	77.30
c=11	0.46	0.53	1.19	237.96	10.14	214.95	**554.30**	4499.80	83.28
c=12	0.43	0.55	0.50	106.46	8.15	230.46	554.92	**4441.20**	89.64

Table 8 Values of c preferred by each cluster validity index for six data sets.

Data sets	c_{opt}	v_{PC}	v_{PE}	v_{XB}	v_K	v_{CWB}	$v_P^{0.1}$	$v_P^{0.01}$	$v_P^{0.001}$	v_{RI}
X30	3	**3**	**3**	**3**	2	**3**	**3**	**3**	**3**	**3**
Bensaid	3	**3**	2	**3**	**3**	6	**3**	7	7	**3**
AD-2	2	**2**	**2**	**2**	**2**	4	4	17	17	**2**
AD-9	9	2	2	**9**	**9**	**9**	**9**	13	13	**9**
AD-3D	4	2	2	2	2	3	**4**	**4**	**4**	**4**
Starfield	9	2	2	**9**	3	**9**	**9**	**9**	**9**	**9**
Iris	2	**2**	**2**	**2**	**2**	3	**2**	11	12	**2**

that can be assigned to 9 clusters by reasonably optimal partitions [15]. Table 6 lists the results of validity indexes for $c = 2, 3, \ldots, c_{max} = 10$. Of the indexes considered, v_{XB}, v_{CWB}, $v_P^{0.1}$, $v_P^{0.01}$, $v_P^{0.001}$, and v_{RI} correctly specified the optimal number of clusters as 9. v_{PC} and v_{PE} considered two clusters to be a natural structure, and v_K points to $c = 3$ clusters as the optimal partition.

Iris data set is known to have 3 clusters; however, two of the clusters are highly overlapped [4]. In view of the geometric structure of Iris data, discussed previously by Pal and Bezdek [4], we took the optimal number of clusters for this data set as 2. Table 7 shows the validation results of each index for $c = 2, 3, \ldots, c_{max} = \sqrt{n} \approx 12$. The optimal $c = 2$ is identified by v_{PC}, v_{PE}, v_{XB}, $v_P^{0.1}$, and v_{RI}. v_{CWB} pointed $c = 3$ as the optimal number of clusters. In contrast, $v_P^{0.01}$ and $v_P^{0.001}$ both failed to detect the optimal number of clusters, and also show the tendency to decrease with increasing c. Once again, the sensitivity of v_P to the choice of the level set causes this index to fail.

Table 8 summarizes the results obtained when each validity index was applied to the seven data sets. The column c_{opt} gives the optimal number of clusters for each data set, and the other columns show the optimal cluster numbers obtained using each index. The proposed index v_{RI} is the only index that correctly recognizes the number of clusters for all data sets. v_P also shows the superior performance to other indexes at $\mu \in \{0.1, 0.2, \ldots, 1.0\}$; however, its performance is dependent on the choice of the level set, $\{\mu\}$. For $\mu \in \{0.01, 0.02, \ldots, 1.0\}$ and $\mu \in \{0.001, 0.002, \ldots, 1.0\}$, $v_P^{0.01}$ and $v_P^{0.001}$ show the tendency to decrease with increasing c, and hence incorrectly identify the optimal c for four of the seven data sets. Although $v_P^{0.1}$ showed less decreasing tendency than $v_P^{0.01}$ and $v_P^{0.001}$, it was also problematic when applied to the AD-2 data set. v_{XB} correctly identifies the optimal c in all data sets except AD-3D. v_{PC} and v_{PE} incorrectly identify the optimal as $c = 2$ for the AD-9, AD-3D, Starfield data sets. The index v_K fails to recognize c_{opt} in the X30, AD-3D, Starfield data sets, while v_{CWB} shows correct validation results in three of the seven data sets.

5. Discussion and Conclusions

In this paper, the problems of conventional validity indexes are reviewed and two of the shortcomings of the validity index of Kim et al. [15] (v_P) are examined. A new cluster validity index for the FCM algorithm is proposed. This validity index is defined as the average value of the relative intersections of all possible pairs of fuzzy clusters in the system. It computes the overlap of each pair of fuzzy clusters by considering the intersection of each data point in the overlap. The optimal number of clusters is obtained by minimizing the validity index. Finally, the performance of the proposed validity index was tested by applying it to well known data sets and comparing the results with those obtained using several other validity indexes. The results indicate that the proposed validity index is very reliable.

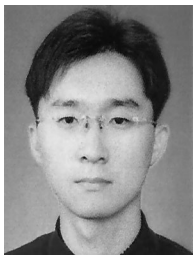
However, the proposed approach is not without drawbacks. Like every other validity index for FCM, the proposed index depends on results obtained using the FCM algorithm. If FCM falls into local optima, the evaluation of validity indexes is useless. The data sets considered here were limited to hyper-spherical shapes because FCM uses a centroid prototype. Since the proposed validity index does not explicitly rely on the centroid prototype, we plan to apply the proposed validity index to other fuzzy clustering algorithms in the future.

Acknowledgments

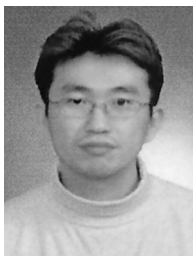
This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc).

References

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, 1991.
[2] J.C. Bezdek, "Numerical taxonomy with fuzzy sets," J. Math. Biol., vol.1, pp.57–71, 1974.
[3] J.C. Bezdek, "Cluster validity with fuzzy sets," J. Cybern., vol.3, pp.58–72, 1974.
[4] 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, 1995.
[5] 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, 1991.
[6] S.H. Kwon, "Cluster validity index for fuzzy clustering," Electron. Lett., vol.34, no.22, pp.2176–2177, 1998.
[7] M.R. Razaee, B.P.F. Lelieveldt, and J.H.C. Reiber, "A new cluster validity index for the fuzzy c-means," Pattern Recognit. Lett., vol.19, pp.237–246, 1998.
[8] A.K. Jain and R.C. Dubes, Algorithms for Clustering, Prentice-Hall, NJ, 1998.
[9] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data clustering: A review," ACM Comput. Surv., vol.31, no.3, pp.264–323, 1999.
[10] H. Lee-Kwang and K.M. Lee, "Fuzzy hypergraph and fuzzy partition," IEEE Trans. Syst. Man Cybern. B, Cybern., vol.25, no.2, pp.196–201, 1995.
[11] H. Lee-Kwang, K.A. Seong, and K.M. Lee, "Hierarchical partition of nonstructured concurrent systems," IEEE Trans. Syst. Man Cybern. B, Cybern., vol.27, no.1, pp.105–108, 1997.
[12] J.C. Bezdek, J. Keller, R. Krisnapuram, and N.R. Pal, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Kluwer Academy Publishers, Boston, 1999.
[13] A.M. Bensaid et al., "Validity-guided (re)clustering with applications to image segmentation," IEEE Trans. Fuzzy Syst., vol.4, no.2, pp.112–123, 1996.
[14] J.C. Bezdek and N.R. Pal, "Some new indexes of cluster validity," IEEE Trans. Syst. Man Cybern. B, Cybern., vol.28, no.3, pp.301–315, 1998.
[15] D.W. Kim, K.H. Lee, and D. Lee, "Fuzzy cluster validation index based on inter-cluster proximity," Pattern Recognit. Lett., vol.24, no.15, pp.2561–2574, 2003.
[16] D. Dumitrescue, B. Lazzerini, and L.C. Jain, Fuzzy Sets and Their Application to Clustering and Training, CRC Press, 2000.



Dae-Won Kim received the M.S. and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology (KAIST), Korea in 1999 and 2004. Currently, he is a post-doctoral fellow at CHUNG Moon Soul Center for Bioinformation and Bioelectronics, Department of BioSystems, KAIST. His research interests include data mining, bioinformatics, and machine learning.



Young-il Kim received the M.S. degree in computer science from Korea Advanced Institute of Science and Technology (KAIST), Korea in 1997. He is now pursuing a Ph.D. degree at the Department of Computer Science, KAIST. His research interests include artificial intelligence and clustering.

Doheon Lee received the B. S., M. S., and Ph. D. degrees in computer science from KAIST, Daejeon, Korea, in 1990, 1992, and 1995, respectively. He has conducted visiting researches in University of Texas, Austin, and National Institutes of Health, Bethesda, in 1999 and 2002, respectively. He is now an Associate Professor in Department of BioSystems, KAIST. His research interests include bio-data mining, bio-system modeling, and bioinformatics. Dr. Lee is the Editor-in-Chief for BioSystems Review, and an Associate Editor for ACM Transactions on Internet Technology.

Kwang Hyung Lee received D.E.A and Dr.Ing. degrees from the Department of Computer Science, INSA de Lyon University, France, in 1982 and 1985, respectively, and the Dr.Etat degree from the Department of Computer Science, INSA de Lyon University, France, in 1988. He is a professor in Department of Computer Science and Department of BioSystems, and a chair professor of Mirae Corporation and the dean of international affair office at KAIST. His research interests include fuzzy systems, artificial intelligence, and bioinformatics.