Evaluation measures for cluster ensembles based on a fuzzy generalized Rand index

Chiu-Chereng Yeh, Miin-Shen Yang

A R T I C L E   I N F O

Article history:
Received 10 October 2014
Received in revised form 25 March 2017
Accepted 26 March 2017
Available online 6 April 2017

Keywords:
Similarity
Rand index
Crisp partition
Fuzzy partition
Fuzzy clustering
Cluster ensemble
Relation matrix
Consensus matrix

A B S T R A C T

Cluster ensemble has become a general technique for combining multiple clustering partitions. There are various cluster ensemble methods to be used in real applications. Recently, Zhang et al. (2012) considered a generalized adjusted Rand index (GARI) for cluster ensembles by using a consensus matrix to evaluate ARI values. However, Zhang’s method for cluster ensembles cannot treat the cases in fuzzy partitions and fuzzy cluster ensembles. In this paper we propose evaluation measures for cluster ensembles based on the proposed fuzzy generalized Rand index (FGRI). We first use a graph and relation matrices to convert a membership matrix into a sign relation matrix, and have the trace of matrix multiplication to calculate similarity measures. We then use the FGRI to broaden the scope of the RI for considering other scenarios so that it can treat the following situations: (1) between a fuzzy cluster ensemble and a crisp partition, (2) between a fuzzy cluster ensemble and a cluster ensemble, (3) between a fuzzy cluster ensemble and a fuzzy partition, (4) between two fuzzy cluster ensembles, and (5) between two different object data sets with the same cardinal number and the same partition method. Finally, numerical comparisons and experimental results are used to demonstrate the key properties, rationality, and practicality of the proposed method.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Cluster analysis is important in data science. Clustering is a method for finding clusters in a data set characterized by the greatest similarity within the same cluster and the greatest dissimilarity between different clusters. Clustering algorithms are useful tools for cluster analysis [22,24,36]. There are various clustering algorithms that have been proposed in the literature, but there is less single one being able to work well for different data sets. Combining those partitions from different clustering algorithms by using cluster ensemble becomes a useful clustering framework. This technique is generally called cluster ensemble. Cluster ensemble has been widely used in many application areas, such as machine learning [6,11,15], bioinformatics [10,20,26], image segmentation [47], data mining [32,35], pattern recognition [12,13].

Strehl and Ghosh [31] first proposed cluster ensemble with three effective methods for combining multiple partitions on collected data sets to obtain high-quality combiners (ensembles). One of these methods involves using the relation matrix of all crisp partitions for defining the representative matrix of the ensemble results. Monti et al. [26] called this type of representative matrix a consensus matrix. This method is both rational and simple. Subsequently, there are many cluster ensemble methods proposed in the literature [3,4,16,34,37]. In general, one of the most popular approaches for combining multiple partitions among cluster ensemble techniques is to construct a consensus matrix. In this sense, evaluating the consistency between consensus matrices in cluster ensembles is important. However, there is less evaluation measure considered in the literature. In this paper we make an effort to advance evaluation measures for cluster ensembles.

In general, Jaccard index (JI) [21], Rand index (RI) [29] and adjusted Rand index (ARI) [18] are the most known indices for measuring the similarity between crisp partitions, and has been widely used in various areas [9,23,33,44]. However, these indices can be only used for comparing similarity measures between crisp partitions, especially for the reference partitions and these partitions produced by the k-means clustering. In fact, there is less evaluation measure between consensus matrices from cluster ensembles. According to our best knowledge, only the paper of Zhang et al. [48] considered the comparisons between cluster ensembles based on their respective consensus matrices.
Zhang et al. [48] considered a generalized adjusted Rand index (ARI) for cluster ensembles by using a consensus matrix to compute the ARI values for the following two cases: (i) a cluster ensemble and a crisp partition, and (ii) two cluster ensembles. However, their method [48] cannot treat the cases in fuzzy partitions and fuzzy cluster ensembles. We know that fuzzy clustering algorithms had been widely studied and applied in various areas [see [5,27,38–41]] so that evaluation measures for fuzzy partitions and fuzzy cluster ensembles are important. Therefore, extending evaluation measures from partitions and cluster ensembles to fuzzy partitions and fuzzy cluster ensembles is also one of our main purposes in this paper, and it is supposed to be important and also the first work in the literature. We follow our recent work in fuzzy generalized Rand index (Yang and Yeh [42]), and then propose evaluation measures for cluster ensembles based on the proposed fuzzy generalized Rand index to broaden the evaluation scope to fuzzy situations such as: between a fuzzy cluster ensemble and a crisp partition, between a fuzzy cluster ensemble and a cluster ensemble, between a fuzzy cluster ensemble and a fuzzy partition, between two fuzzy cluster ensembles, and so forth.

The remainder of the paper is organized as follows. In Section 2, we first review Rand index, other related indexes, and cluster ensemble. We then review the method proposed by Zhang et al. [48] where we also give some descriptions about the shortcoming of Zhang’s method. We also review some existing extensions for Rand index. In Section 3, we first present the fuzzy generalized Rand index. We then construct the evaluation measures for cluster ensembles based on the proposed fuzzy generalized Rand index. In Section 4, numerical comparisons and experimental results are used to clarify the rationality and practicality of the proposed method. Finally, conclusions are stated in Section 5.

2. Rand index, other related indexes and cluster ensemble

Assume that there are two crisp partitions $P^{(r)}$, $r=1,2$ in the object data set $O = \{o_1, o_2, \cdots, o_n\}$ with $P^{(r)} = \{S_1^{(r)}, S_2^{(r)}, \cdots, S_{k_r}^{(r)}\}$ where $k_r = \sum_{h=1}^{n} n_{h}^{(r)} = \text{Oand}S_{h}^{(r)} \cap S_{h} \neq \emptyset$ for all $h \neq h'$, and $k_r$ denotes the number of clusters of the partition $P^{(r)}$. Let $a$ indicate the number of the pairs of $o_i$ and $o_j$ belonging to the same cluster in $P^{(1)}$ and $P^{(2)}$. Let $b$ indicate the number of the pairs of $o_i$ and $o_j$ belonging to the same cluster in $P^{(1)}$ and to different clusters in $P^{(2)}$. Let $c$ indicate the number of the pairs of $o_i$ and $o_j$ belonging to different clusters in $P^{(1)}$ and to the same cluster in $P^{(2)}$. Let $d$ indicate the number of the pairs of $o_i$ and $o_j$ belonging to different clusters in $P^{(1)}$ and $P^{(2)}$. Let $n_{uv}$ represent the number of objects that belong to $S_u^{(1)}$ in the partition $P^{(1)}$ and $S_v^{(2)}$ belong to in the partition $P^{(2)}$.

Let $n_{uv} = \sum_{i=1}^{k_2} n_{uv_i}$, $n_{v} = \sum_{u=1}^{k_1} n_{uv}$, and $\binom{N}{k} = \frac{N!}{k!(N-k)!}$. Then, the Rand index (RI) [29] between the crisp partitions $P^{(1)}$ and $P^{(2)}$ can be defined as follows:

$$\text{RI}(P^{(1)}, P^{(2)}) = \frac{a + d}{a + b + c + d} = \frac{(\binom{n}{2}) + \sum_{u=1}^{k_2} \sum_{v=1}^{k_1} n_{uv}^2 - \frac{1}{2} \sum_{u=1}^{k_2} n_{u}^2 + \frac{1}{2} \sum_{v=1}^{k_1} n_{v}^2}{\binom{n}{2}}$$

(1)

Assume that the crisp partition matrix $M_{C}^{(r)} = \begin{bmatrix} n_{hi}^{(r)} \end{bmatrix}_{k_r \times n}$ of $P^{(r)}$, $r=1,2$, is defined as $n_{hi}^{(r)} = \begin{cases} 1 & \text{if } o_i \in S_{h}^{(r)} \\ 0 & \text{otherwise} \end{cases}$. Let $N = (M_{C}^{(1)})^T(M_{C}^{(2)})^T$, where $(A)^T$ indicates the transpose of the matrix $A$. The notation $\#(S)$ denotes the number of elements in the set $S$. Then $N = [n_{uv}]_{k_1 \times k_2}$, where $n_{uv} = \#(S_u^{(1)} \cap S_v^{(2)})$. Assume that $N$, where $N$ is called as the $k_1 \times k_2$ contingency table as shown in Table 1, is constructed from the generalized hypergeometric distribution and the maximum of $RI$ equals to 1. Then the adjusted Rand index (ARI) [18] between the crisp partitions $P^{(1)}$ and $P^{(2)}$ is defined as

$$\text{ARI}(P^{(1)}, P^{(2)}) = \frac{RI - E(RI)}{\text{max}(RI) - E(RI)} = \frac{\frac{k_1}{2} \sum_{u=1}^{k_2} n_{uv}^2 + \frac{k_2}{2} \sum_{v=1}^{k_1} n_{uv}^2 - \frac{2k_1}{n(n-1)} \sum_{u=1}^{k_1} \binom{n_u}{2} - \frac{2k_2}{n(n-1)} \sum_{v=1}^{k_2} \binom{n_v}{2}}{k_1 - I_{0} - I_{1} - I_{2}}$$

(2)

where $E(RI)$ is the expected value of $RI$ and $\text{max}(RI)$ is the maximum of $RI$. Let

$$I_0 = \sum_{i=1}^{k_1} \binom{n_i}{2}, I_1 = \sum_{i=1}^{k_2} \binom{n_i}{2}, I_2 = \sum_{v=1}^{k_2} \binom{n_v}{2}, I_{3} = \frac{2 - I_0}{n(n-1)}$$

Then

$$\text{ARI}(P^{(1)}, P^{(2)}) = \frac{I_0 - I_3}{\frac{2}{k_1(k_1 + k_2)} - I_{3}} = \frac{a}{\frac{2(b + c)(a + c)}{n(n-1)} - \frac{2(b + c)(a + c)}{n(n-1)}}$$

(3)

Next, we review a consensus matrix. Assume that a crisp partition of the object data set $O = \{o_1, o_2, \cdots, o_n\}$ is $P = \{S_1, S_2, S_3, \cdots, S_k\}$, where $\bigcup_{h=1}^{n} S_{h} = \text{Oand}S_{h} \cap S_{h} \neq \emptyset$ for all $h \neq h'$. Then the relation matrix $R = [r_{ij}]_{n \times n}$ of $O$ is defined as follows:

$$r_{ij} = \begin{cases} 1 & \text{if } o_i \text{ and } o_j \text{ belong to the same cluster in } P \\ 0 & \text{otherwise} \end{cases}$$

Let the co-association matrix $A$ for $P$ be $A = R - I = [a_{ij}]_{n \times n}$, where $I$ is an $n \times n$ identity matrix. Because of various consideration, we assume that the same object data set $O$ has the $q'$ crisp partitions $(1)^{P}, (2)^{P}, \cdots, (q')^P$. We know that a cluster ensemble is used to combine these $q'$ crisp partitions $(1)^P, (2)^P, \cdots, (q')^P$ into a useful clustering framework. Let the co-association matrix for $(w)^P$ be defined as $(w)^A = [w_{ij}]_{n \times n}$, where $w = 1, \cdots, q'$. A consensus matrix of cluster ensemble can be constructed as

<table>
<thead>
<tr>
<th>Table 1</th>
<th>The contingency table $N$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>$S_1^{(2)}$</td>
</tr>
<tr>
<td>$S_1^{(1)}$</td>
<td>$n_{11}$</td>
</tr>
<tr>
<td>$S_2^{(1)}$</td>
<td>$n_{21}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$S_{k_1}^{(1)}$</td>
<td>$n_{k_11}$</td>
</tr>
<tr>
<td>$\sum$</td>
<td>$n_1$</td>
</tr>
</tbody>
</table>
$C = [c_{ij}]_{n \times n}$ where $c_{ij} = \frac{1}{d} \sum_{w=1}^{d} a_{wj}$. Zhang et al. [48] proposed an evaluation measure by using the entries of the consensus matrix to find ARI values. The method of Zhang et al. [48] is described as follows.

(i) Let $P = \{S_1, S_2, \ldots, S_k\}$ be the crisp partition of the object data set $O = \{o_1, o_2, \ldots, o_n\}$, where $\#(S_i) = n_h$ for $h = 1, \ldots, k$. Assume that there are $q$ crisp partitions $(P_1^q, P_2^q, \ldots, P_q^q)$ for the object data set $O$. Let $E$ be a cluster ensemble from the $q$ crisp partitions $(P_1^q, P_2^q, \ldots, P_q^q)$, where $E$ has its consensus matrix $C$ with
\[
C = [c_{ij}]_{n \times n}
\]
where $c_{ij} = \frac{1}{q} \sum_{w=1}^{q} a_{wj}$. Let $z_0 = \sum_{h=1}^{k} \sum_{i \neq j} c_{ij} = \sum_{i \neq j} z_2$.
\[
z_2 = \sum_{h=1}^{k} \left( \frac{n_h}{2} \right)
\]
and $z_3 = \frac{2z_1z_2}{n(n-1)}$. Then $ARI$ between the cluster ensemble $E$ and the crisp partition $P$ is defined as
\[
ARI_{mp}(E, P) = \frac{z_0 - z_3}{z_2 + z_3 - z_3}
\]
(ii) Assume that there are two cluster ensembles, written as $E_1$ and $E_2$, for the object data set $O$ and their consensus matrices are $C(1) = [c_{ij}^{(1)}]_{n \times n}$ and $C(2) = [c_{ij}^{(2)}]_{n \times n}$ respectively. Let $t_0 = \sum_{i \neq j} c_{ij}^{(2)}$, $t_1 = \sum_{i \neq j} c_{ij}^{(2)}$, $t_2 = \frac{t_1}{2}$, and $t_3 = \frac{2z_1z_2}{n(n-1)}$. Then, $ARI$ between the two cluster ensembles $E_1$ and $E_2$ is defined as
\[
ARI_{mm}(E_1, E_2) = \frac{t_0 - t_3}{t_2 + t_3 - t_3}
\]
We see that Zhang et al. [48] substituted $z_0$ for $l_0$, $z_1$ for $l_1$, and $z_2$ for $l_2$ in Eq. (3), and then used Eq. (4) to evaluate $ARI$ between the cluster ensemble and the crisp partition. Similarly, Zhang et al. [48] substituted $t_0$ for $bu$, $t_1$ for $l_1$, and $t_2$ for $l_2$, and then used Eq. (5) to evaluate $ARI$ between two cluster ensembles. Using the same substitution way, successfully computing other similarities, such as $RI$ or the jaccard index $[J]$, is also available. However, a consensus matrix is defined as an average matrix between $q'$ co-association matrices. This definition enables decimals to appear in the number of pairs of two different objects belonging to the same cluster. The original definition of both $ARI$ and $Rand$ contains the restriction that $S(h)^r \cap S(h')^r = \emptyset$ for all $h \neq h'$. Therefore, for all instances of $i$ and $j$, the only possible states for a paired relationship between $o_i$ and $o_j$ are yes or no, i.e., 0 or 1. In this sense, Zhang et al. [48] violates the original definition of $ARI$ or $Rand$. Thus, the method proposed by Zhang et al. [48] is necessary to be improved. To improve the shortcoming of Zhang et al. [48], we propose an evaluation method by using the fuzzy generalized Rand index in the next section. On the other hand, our proposed method can also treat the evaluation for cluster ensembles in fuzzy cases.

Before we continue to the next section, we need to make some notices. According to Eq. (2), $ARI$ is established in the contingency table (i.e., Table 1), but the condition of the generalized hypergeometric distribution needs to be satisfied, and the index defined by the expected value $E(\sum_{i}^{d} \sum_{j} \left( \frac{n_{ij}}{2} \right) )$ and max($RI$) = 1 is also employed. However, the $ARI$ computation of Zhang et al. [48], such as $ARI_{mp}$ and $ARI_{mm}$, is not the same as the original $ARI$ computation under the assumption of hypergeometric distributions. Zhang et al. [48] considered his construction by replacing the $ARI$ computation with a consensus matrix of cluster ensemble on an averaging approach without any probabilistic assumption. Similarly, the fuzzy relation between the object $o_i$ and the cluster $S_h$ is defined by a partition membership $m_{hi}$, where $m_{hi} \in [0, 1]$ and $\sum_{h=1}^{k} m_{hi} = 1$, so a nonparametric approach without any probabilistic assumption should be considered for fuzzy partition cases. In this paper we are interested not only in the cluster ensemble problem of crisp partitions, but also in the cluster ensemble problem of fuzzy partitions. Therefore, to fully decode the information and facilitate a rigorous analysis, we did not adopt $ARI$, but use $RI$ to extensions of evaluation measures for cluster ensembles. This is because $RI$ does not require any probability assumption.

Next, we review several fuzzy extensions for the Rand index. Let $M^{(r)}_h = \left[ m_{hi}^{(r)} \right]_{k \times n}$, $r = 1, 2$, be a fuzzy partition matrix between the set $O = \{ o_1, o_2, \ldots, o_n \}$ and the fuzzy partition $P^{(r)} = \{ S^{(r)}_1, S^{(r)}_2, \ldots, S^{(r)}_k \}$, where a membership $m_{hi}^{(r)}$ in [0, 1] indicates the degree in which $o_i$ belongs to the cluster $S^{(r)}_h$. Campello [8] used set-theoretic to extend $RI/\text{FC}R$ for short. Let $G^{(r)} = \left[ g_{ij}^{(r)} \right]$ where $g_{ij}^{(r)}$ is the maximum of the minimum membership of $(o_i, o_j)$ belonging to the same cluster in $P^{(r)}$, and $W^{(r)} = \left[ w_{ij}^{(r)} \right]$ where $w_{ij}^{(r)}$ is the maximum of the minimum membership of $(o_i, o_j)$ belonging to different clusters in $P^{(r)}$. Campello [8] proposed the following equations:
\[
\begin{align*}
\alpha &= \sum_{i=1}^{n} \sum_{j=1}^{n} \min(g_{ij}^{(1)}, g_{ij}^{(2)}), b &= \sum_{i=1}^{n} \sum_{j=1}^{n} \min(g_{ij}^{(1)}, w_{ij}^{(2)}), c &= \sum_{i=1}^{n} \sum_{j=1}^{n} \min(v_{ij}^{(1)}, v_{ij}^{(2)}), d &= \sum_{i=1}^{n} \sum_{j=1}^{n} \min(v_{ij}^{(1)}, v_{ij}^{(2)}),
\end{align*}
\]
and
\[
FRC(M^{(1)}, M^{(2)}) = FRC(P^{(1)}, P^{(2)}) = \frac{\alpha b}{2 \alpha d}
\]
Brouwer [7] used a vector to deal with the membership of objects $o_i$ belonging to each cluster (FRB for short). Let matrix $U^{(r)} = [u_{hi}^{(r)}]_{k \times n} = \left[ m_{hi}^{(r)} \right]_{k \times n}$ where $m_{hi}^{(r)}$ is the length of $i$th column vector of the fuzzy partition matrix $M^{(r)}_h$, $i = 1, 2, \ldots, n$, $r = 1, 2$. Then, the bonding matrix is defined as $B^{(r)} = \left( U^{(r)} \right)^T \left( U^{(r)} \right)$, and Brouwer [7] gave the following equations:
\[
\begin{align*}
\alpha &= g(B^{(1)} \wedge B^{(2)})^T, b &= f(B^{(1)} \wedge (-B^{(2)}))^T, c &= f((-B^{(1)}) \wedge B^{(2)})^T), d &= f((-B^{(1)}) \wedge (-B^{(2)}))^T,
\end{align*}
\]
where $f(Z) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} z_{ij}$, $g(Z) = f(Z) - \frac{Z}{2}$, $\neg B^{(r)} = I - B^{(r)}$, $r = 1, 2$, and $J$ is a $n \times n$ matrix with all entries 1s. Anderson et al. [1] recommended replacing the term $N = \left( M^{(1)} \right)^T M^{(2)}$ in $RI$ with $N^* = \varphi \cdot N = \left[ n_{ij}^* \right]_{k \times k}$ (FCR for short), where
\[
\varphi = \left\{ \begin{array}{ll}
1 & \text{if} M^{(1)}, M^{(2)} \text{are crisp, fuzzy or probabilistic,} \\
\sum_{i} \sum_{j} n_{ij} & \text{if} M^{(1)} \text{and/or} M^{(2)} \text{are possibilistic.}
\end{array} \right.
\]
Hullermeier et al. [19] considered the Rand index as a distance function (FRH for short). Let $E_{M^{(r)}}^{(c)}(o_i, o_j) = 1 - \frac{1}{2} \sum_{h=1}^{k} | m_{hi}^{(r)} - m_{hj}^{(r)} | = 1 - ||m_{hi} - m_{hj}||_1$, $d(M^{(1)}, M^{(2)}) = \frac{1}{2} \sum_{h=1}^{k} \sum_{i, j} \left| E_{M^{(r)}}^{(c)}(o_i, o_j) - E_{M^{(r)}}^{(c)}(o_i, o_j) \right|$. Then, the fuzzy extension of the Rand index is equal to $\varphi$.

3. The proposed evaluation measure for cluster ensembles

In this section we propose an evaluation measure by using a newly proposed fuzzy generalized Rand index. To improve the shortcoming that causes a decimal to appear in $n_{hi}$ in the method
proposed by Zhang et al. [48], we will not use a consensus matrix for cluster ensembles. This is because entries of a consensus matrix may appear decimal points. We note that there are several fuzzy extensions of RI in the literature (see [17,8,19]). However, our recently proposed fuzzy generalized Rand index [42] is well used as an evaluation measure for cluster ensembles. We first use a membership matrix to find a similarity measure for cluster ensembles and then extend it to a signature relation matrix for replacing the consensus matrix. Finally, we use the trace of matrix (from the result of two sign relation matrix multiplication) to compute FGRJ (fuzzy generalized Rand index). For clarity of description, we divide the following parts to discuss.

(i) Let the crisp partition matrices of $(1)p, (2)p, \ldots, (q)p$ in the same object data set $O = \{o_1, o_2, \ldots, o_n\}$ be $(1)M_{C_{(1)}}, (2)M_{C_{(2)}}, \ldots, (q)M_{C_{(q)}}$, respectively. Let $E$ stand for the cluster ensemble that consists of $(1)p, (2)p, \ldots, (q)p$. The number of clusters of $(w)p$ is denoted by $k_w$, $w = 1, 2, \ldots, q$. Let the membership matrix $M = [m_{hi}]_{k \times n}$ stand for the cluster ensemble $E$, where $k' = \max(k_1, k_2, \ldots, k_q)$ and $m_{hi} = \{1 \over q} \sum_{w=1}^{q} (w)m_{hi}$. Let $M^{(r)} = [m^{(r)}_{hi}]_{k' \times n}$ be the membership matrix of the cluster ensemble $E_r$, where $r = 1, 2, \ldots, n$. We propose an evaluation method for these membership matrices $M^{(r)} = [m^{(r)}_{hi}]_{k' \times n}$ by extensions of RI. However, RI cannot directly be used for defining these membership matrices. This is because $RI$ is originally defined by binary logic (i.e. binary value 0 or 1) for crisp partition matrices, and the entries of these membership matrices $M^{(r)} = [m^{(r)}_{hi}]_{k' \times n}$ may appear decimal points (because $m^{(r)}_{hi} \in [0, 1]$). In fact, all of $RI[29], ARI[18], IJ[21]$ and most other indices [2,12,17] used a ratio to define a similarity measure between two partitions. It is seen that the value of a ratio will not be changed when the units of measurements decrease, so we consider splitting each object $o_i$ in the object data set $O = \{o_1, o_2, \ldots, o_n\}$ into $t$ equal smaller objects $o_{i1}, o_{i2}, \ldots, o_{it}$. This splitting idea seems to take each object in the object data set as a combination of these $t$ smaller objects. In other words, we view each object as an assembly of several smaller objects to the condition that the only possible states for the relations between them are “belong to” or “not belong to” with only binary value 0 or 1. On the other hand, multiplying the numerator and denominator of a ratio by a nonzero number does not alter the actual value of the ratio. That is, minimizing a unit does not change the ratio; therefore, splitting does not change $RI$ or $JI$. In this sense, we can extend $RI$ or $JI$ for treating these membership matrices $M^{(r)} = [m^{(r)}_{hi}]_{k' \times n}$ with $m^{(r)}_{hi} \in [0, 1]$. In general, both $RI$ and $JI$ yield similar results in the most examples. Of course, we only consider the most common method, $RI$, for extension.

Let each $o_i$ be divided into $t$ equal elements $o_{i1}, o_{i2}, \ldots, o_{it}$, where $t$ is a fixed number such that all $t \times m^{(r)}_{hi}$, $r = 1, 2, \ldots, n$, become integers, and so the object data set $O = \{o_1, o_2, \ldots, o_n\}$ is extended to be the object data set $O = \{o_{11}, o_{12}, \ldots, o_{i1}, o_{i2}, \ldots, o_{it}, o_{11}, o_{12}, \ldots, o_{n1}, o_{n2}, \ldots, o_{nt}\}$. Therefore, the number of elements of $O = nt$ (i.e. #($O_g$) = $nt$). After splitting each object $o_i$, there are $t \times m^{(r)}_{hi}$ small objects belonging to the cluster $S^{(r)}_{hi}$ of cluster ensemble $E_r$ in the set $[o_{11}, o_{12}, \ldots, o_{it}]$. Note that the variable $r$ plays a vital role. This variable $r$ not only keeps the extension in a binary form (i.e. “belong to” or “not belong to”), but also prevents loss of data caused by using other methods (e.g., bounded difference, algebraic product, or the minimum) to evaluate fuzzy intersections. The variable $t$ also preserves the original intentions of the scholars who defined similarity measures. For example, in the $RI$ proposed by Rand [29] and the $ARI$ proposed by Hubert and Arabie [18], each $n_{am}$ is a positive integer or zero, and $S^{(r)}_{h_i} \cap S^{(r)}_{h_j} = \emptyset$ for all $h \neq h'$.

Let $o_{im} \in O_E$ with a “$+$” line, say “$+$” edge if the pair of $o_{im}$ and $o_{ij}$ in $O_E$ belongs to the same cluster in the cluster ensemble $E$, and let $o_{im} \in O_E$ with a “$-$” line, say “$-$” edge if the pair of $o_{im}$ and $o_{ij}$ in $O_E$ belongs to different clusters in the cluster ensemble $E$. Then, the sign relation matrix $m^{(r)}_{hi} = |m^{(r)}_{hi}|_{k \times n}$ for the membership matrix $M^{(r)}$ is defined as follows.

\[
m^{(r)}_{hi} = \begin{cases} 
0 & \text{if } i = j \\
n & \text{if } o_{im} \text{ and } o_{ij} \text{ joined with } + \text{ edge in cluster ensemble } E, \text{ where } i \neq j \\
-1 & \text{if } o_{im} \text{ and } o_{ij} \text{ joined with } - \text{ edge in cluster ensemble } E, \text{ where } i \neq j
\end{cases}
\]

(6)

Note that for all $u$, $v$, $r_{uv} = 0$ if $i = j$. This is because any pair $(o, o_j)$ in the object data set $O$ cannot draw a simple edge. Let $Y = (R^{(1)}(Y), R^{(2)}(Y))$. Then, we define the fuzzy generalized Rand index (FGRJ) between two cluster ensembles $E_1$ and $E_2$ as follows:

\[
FGRJ(E_1, E_2) = FGRJ(M^{(1)}), M^{(2)}) = \frac{1}{2} + \frac{1}{2} \frac{1}{n} (Tr(Y) - Tr(C))
\]

(7)

Where $Tr(Y)$ denotes the trace of $Y$.

In addition, we also consider how to use a sign relation matrix to individually solve $a,b,c,$ and $d$ in Eq. (1) for the different similarity indices (for example, $JI$ or $ARI$, etc.) proposed by various scholars. For any $a_{im}, o_{im} \in O_E$, let $R^{(r)} = [r_{ij}]_{nt \times nt}$ and $R^{(r)} = [-r_{ij}]_{nt \times nt}$ be defined as follows:

\[
\begin{align*}
&\quad r^{(r)}_{im} = \begin{cases} 
0 & \text{if } i = j \\
1 & \text{if } o_{im} \text{ and } o_{ij} \text{ joined with } + \text{ edge in cluster ensemble } E, \text{ where } i \neq j \\
0 & \text{if } o_{im} \text{ and } o_{ij} \text{ joined with } - \text{ edge in cluster ensemble } E, \text{ where } i \neq j
\end{cases} \\
&\quad r^{(r)}_{im} = \begin{cases} 
0 & \text{if } i = j \\
0 & \text{if } o_{im} \text{ and } o_{ij} \text{ joined with } + \text{ edge in cluster ensemble } E, \text{ where } i \neq j \\
-1 & \text{if } o_{im} \text{ and } o_{ij} \text{ joined with } - \text{ edge in cluster ensemble } E, \text{ where } i \neq j
\end{cases}
\end{align*}
\]

(8)

\[
\begin{align*}
&\quad a = \frac{1}{2} | Tr(R^{(1)}) | Tr(R^{(2)}) | \\
&\quad b = \frac{1}{2} | Tr(R^{(1)}) | Tr(R^{(2)}) | \\
&\quad c = \frac{1}{2} | Tr([-R^{(1)}) | Tr([-R^{(2)}) | \\
&\quad d = \frac{1}{2} | Tr([-R^{(1)}) | Tr([-R^{(2)}) |
\end{align*}
\]

(9)

(10)

(11)

(ii) Of course, we can also evaluate the similarity between a cluster ensemble and a crisp partition. Assume $M_{c} = [m_{hi}]_{k \times n}$ is the crisp partition matrix of $P = \{S_1, S_2, \ldots, S_k\}$ in the object data set $O = \{o_1, o_2, \ldots, o_n\}$ and $k$ is the number of clusters of $P$. Let the membership matrix $M = [m_{hi}]_{k \times n}$ represent the cluster ensemble $E$ that consists of $(1)p, (2)p, \ldots, (q)p$. Let $M^{(1)} = M$ and $M^{(2)} = M_{c}$. Then, we use the variable $t$ and Eq. (6) to convert $M$ and $M_{c}$ into sign relation matrices, followed by using Eq. (7) to compute $FGRJ$ between a cluster ensemble $E$ and a crisp partition $P$ (i.e. $FGRJ(E, P) = FGRJ(M^{(1)}, M_{c})$).

The process of the proposed method is shown in Fig. 1.

From Fig. 1, the proposed method can also evaluate $RI$ between a cluster ensemble and a fuzzy partition. Assume $M_{F} = [m_{hi}]_{k \times n}$ is the fuzzy partition matrix between the object data set $O$ and the partition $P$. Let the membership matrix $M = [m_{hi}]_{k \times n}$ represent the cluster ensemble $E$ that consists of $(1)p, (2)p, \ldots, (q)p$, where $(w)p, w = 1, 2, \ldots, q$ are crisp partitions in the object data set $O$. A
suitable value for $t$ is then selected to ensure that all entries of $M_F$ and $M'$ are positive integers or zero. We next transform the matrices $M_F$ and $M'$ into sign relation matrices on the step 2 of Fig. 1. We obtain $FGRI(E, P_f) = FGRI(M', M_F)$ on the step 3 of Fig. 1.

(iii) For further generalizations by membership matrices to compute the similarity, the following several cases are also incorporated. Assume that the fuzzy cluster ensemble (be written as $E_F$) is consisted of fuzzy partitions $(1)P_1, (2)P_2, \cdots, (q)P_q$ and their fuzzy partition matrices are $(1)M_F = \left(\begin{array}{c} m_{11}^{(1)} \end{array} \right)_{k \times n}$, $(2)M_F = \left(\begin{array}{c} m_{11}^{(2)} \end{array} \right)_{k \times n}$, $\cdots$, $(q)M_F = \left(\begin{array}{c} m_{11}^{(q)} \end{array} \right)_{k \times n}$, respectively. Let the fuzzy membership matrix $M_F = \left[ \sum_{w=1}^{q} (w)M^{(w)}_{hi} \right]_{k \times n}$ stand for the fuzzy cluster ensemble $E_F$, where $k = \max\{k_1, k_2, \cdots, k_q\}$. Then, we can find out the similarities between the fuzzy cluster ensemble $E_F$ and other partitions by FGRI. For example, assume the fuzzy partition matrix of the fuzzy partition $P_i$ in $E$ is $M_F$, we select a suitable value for $t$ to ensure that all entries of $M_F$ and $M'_F$ are positive integers or zero. Then, we can compute $FGRI(E, M'_F) = FGRI(M'_F, M_F)$ by the process of Fig. 1. Of course, we select a suitable value for $t$ to ensure that all entries of $M_F$ are positive integers or zero if the partition matrix is a crisp partition $M_C$. Then, we can evaluate $FGRI(E, P) = FGRI(M_F, M_C)$. And, let the fuzzy membership matrices of two fuzzy cluster ensembles $E^{(1)}_F$ and $E^{(2)}_F$ are $M^{(1)}_F$ and $M^{(2)}_F$, respectively. Therefore, we can also evaluate the Rand index between two cluster ensembles, i.e., $FGRI(E^{(1)}_F, E^{(2)}_F) = FGRI(M^{(1)}_F, M^{(2)}_F)$.

From above discussion, using a sign relation matrix to compute a similarity of cluster ensembles not only corrects the mentioned irrational results of Zhang et al. [48], but also enables a cluster ensemble (or a fuzzy cluster ensemble) to compare the similarity with other partitions that contain a crisp partition, a cluster ensemble, a fuzzy partition, a fuzzy cluster ensemble. This is because the proposed method preserves the only possible states for the relationship between a small object and the cluster are “belong to” or “not belong to”, but Zhang et al. [48] cannot preserve this property. More different cases can be computed by the proposed method as shown in Fig. 2.

Finally, we mention that, although our extended sign relation matrix has the restriction that it requires the existence of a factor (i.e. the variable $t$) for turning all membership degrees into integers

![Fig. 2. Different cases of FGRI.](image)

![Fig. 1. The process of our proposed method.](image)

**Table 2**

Cluster ensembles and their crisp partitions.

| Cluster ensemble $E$, crisp partition $(w)P^{(r)}$ of $E_i$ |
|-----------------|---------------------------------|
| $E_1$           | $(1)P^{(1)} = \{(01, 02, 03), (04, 05)\}$ |
|                 | $(2)P^{(1)} = \{(01, 02, 03), (04, 05)\}$ |
| $E_2$           | $(1)P^{(2)} = \{(01, 02, 03), (04, 05)\}$ |
|                 | $(2)P^{(2)} = \{(01, 02, 03), (04, 05)\}$ |

with $\sum_{h} m_{hi}^{(r)} = 1$ for all $0 \in O$, this limitation can be overcome.

For example, when we calculate the value of $FGRI$, these $m_{hi}^{(r)}$ can be first rounded to one or two decimal places; thus, the factor-setting problem can be solved. In practical applications, $m_{hi}^{(r)}$ is typically rounded to one or two decimal places because such an approach enables satisfactory similarities to be obtained. When handling ensemble-related problems, if the ensembles $E_1$ and $E_2$ comprise $q_1$ and $q_2$ crisp partitions, respectively, we can choose the factor $t$ as the least common multiple of $q_1$ and $q_2$.

4. Numerical examples and experimental results

In this section, some numerical and real data sets are used to demonstrate the efficiency of the proposed method. We consider $Rl$ as a similarity measure because $Rl$ is the most frequently used and does not require the fulfillment of any probability distribution. In Example 1, we use the proposed method to compute the similarity measures of cluster ensembles. In addition, we also compare it with Zhang et al. [48] method. In Example 2, we use real data to show comparisons of the proposed method with those existing fuzzy extensions of the Rand index [1,7,8,19] and then explain that the proposed method is suitable for fuzzy partitions. In Example 3, crisp partition and fuzzy partitions are used. We then discuss these similarity measures under the following cases: between a fuzzy partition and a crisp partition, between two fuzzy partitions, between a fuzzy cluster ensemble and a crisp partition, between a fuzzy cluster ensemble and a fuzzy partition.

**Example 1.** Assume that the object data set is $O = \{01, 02, 03, 04, 05\}$. Let $E_1$ consist of crisp partitions $(1)P^{(1)}, (2)P^{(1)}$, $(3)P^{(1)}, (4)P^{(1)}, (5)P^{(1)}$. Let $E_2$ consist of crisp partitions $(1)P^{(2)}, (2)P^{(2)}$. For all $r$ and $w$, the crisp partitions $(w)P^{(r)}$ are shown in Table 2.

(1) The method of Zhang et al. [48]

The co-association matrices of $(w)P^{(r)}$ are:

$(1)A^{(1)} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$, etc.
Table 3
RL values.

<table>
<thead>
<tr>
<th>RLj</th>
<th>E1</th>
<th>E2</th>
<th>(3)p(1)</th>
<th>(2)p(1)</th>
<th>(1)p(1)</th>
<th>(4)p(1)</th>
<th>(5)p(1)</th>
<th>(2)p(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>0.68</td>
<td>0.6</td>
<td>0.68</td>
<td>0.6</td>
<td>0.76</td>
<td>0.5</td>
<td>0.6</td>
<td>0.52</td>
</tr>
<tr>
<td>E2</td>
<td>0.6</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 4
FGRI values.

<table>
<thead>
<tr>
<th>FGRI</th>
<th>E1</th>
<th>E2</th>
<th>(3)p(1)</th>
<th>(2)p(1)</th>
<th>(1)p(1)</th>
<th>(4)p(1)</th>
<th>(5)p(1)</th>
<th>(2)p(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>1</td>
<td>0.6500</td>
<td>0.5600</td>
<td>0.4800</td>
<td>0.6560</td>
<td>0.6560</td>
<td>0.7040</td>
<td>0.5440</td>
</tr>
<tr>
<td>E2</td>
<td>0.6500</td>
<td>1</td>
<td>0.5500</td>
<td>0.4500</td>
<td>0.5500</td>
<td>0.5500</td>
<td>0.5500</td>
<td>0.5500</td>
</tr>
</tbody>
</table>

Thus, the RL values from Zhang et al. [48] method (written as RLz) for different situations are shown in Table 3.

(II) The proposed method

\[ C^{(1)} = \begin{bmatrix}
0 & 0.8 & 0.2 & 0.4 & 0.2 \\
0.8 & 0 & 0.4 & 0.2 & 0 \\
0.2 & 0.4 & 0 & 0.4 & 0 \\
0.4 & 0.2 & 0 & 0.6 & 0 \\
0.2 & 0.4 & 0.6 & 0 & 0 \\
\end{bmatrix}. \]

\[ C^{(2)} = \begin{bmatrix}
0 & 0.5 & 0.5 & 0 & 0 \\
0.5 & 1 & 0 & 0 & 0.5 \\
0.5 & 0 & 0 & 0 & 0.5 \\
0 & 0.5 & 0.5 & 0 & 0 \\
\end{bmatrix}. \]

Then, \( M^{(1)} = \begin{bmatrix}
1 & 0.8 & 0.2 & 0.4 & 0.2 \\
0 & 0.8 & 0.2 & 0.4 & 0.2 \\
0 & 0 & 0 & 0.4 & 0.2 \\
\end{bmatrix}. \]

Thus, objects belonging to cluster \( c_{ij}^{(1)} \) within partition \( p^{(1)} \) and belonging to \( c_{ij}^{(2)} \) within partition \( p^{(2)} \) must also be a positive integer or zero. By applying the same logic, \( l_1 = \sum_{k=1}^{k_1} n_{iu} / 2 \) and \( l_2 = \sum_{k=1}^{k_2} n_{iv} / 2 \) must also represent positive integers or zero. However, the cases, such as \( c_{12}^{(1)} = 0.8 \) or \( c_{12}^{(2)} = 0.5 \), are decimal; therefore, attempting to solve \( z_0 = \sum_{h=1/j}^{z} n_{ij} \) for ARI may result in the appearance of decimal points (i.e., \( s_0 \notin N \cup \{0\} \)).

For example, \( s_0 = 1.4 \) when we solve for RLz(\( E_1 \), \( 3p^{(1)} \)) or \( s_0 = 2.5 \) when solving for RLz(\( E_2 \), \( 1p^{(2)} \)). Also, the method for substituting \( l_0 \) with \( l_1 \) and \( l_2 \) with \( l_2 \), proposed by Zhang et al. [48], may appear decimal points. This is the irrational result that \( a, b, c, \) and \( d \) appear decimal points. Furthermore, for all \( r, c^{(r)} \) represent the paired relationship of \( o_i \) and \( o_j \) in the partition \( p^{(l)} \). The original definitions of both the ARI and RL (even if \( \in \)) contain the restriction that \( S_{ij}^{(l)} \cap S_{ij}^{(l)} = \emptyset \) for all \( \neq i \). Therefore, for all instances of \( i \) and \( j \), the only possible states for a paired relationship between \( o_i \) and \( o_j \) are yes or no, i.e. 1 or 0. However, the decimal in \( c_{ij}^{(r)} \) shows that \( c_{ij}^{(r)} \) contains states other than yes or no.

On the other hand, if two cluster ensembles are assumed to be identical, then using the method proposed by Zhang et al. [48] results in RL that is not equal to 1. For example, results such as RLz(\( E_1 \), \( 3p^{(1)} \)) = 0.68 or RLz(\( E_2 \), \( 2p^{(2)} \)) = 0.7 are irrational because cluster ensembles \( E_r \), \( r = 1, 2 \), are formed based on crisp partitions. Logically, RL should equal 1. In this sense, the method proposed by Zhang et al. [48] has the shortcoming. Our proposed method in this paper does not exhibit these limitations. Within the extended object data set \( O = \{ o_{1,1}, o_{1,2}, \ldots, o_{1,10}, o_{2,1}, o_{2,2}, \ldots, o_{2,10}, \ldots, o_{5,10}, \ldots, o_{9,10}, \ldots \} \), the only possible states for the paired relationships between \( o_i \) and \( o_j \) are yes or no (i.e., in the extended sign relation matrix, 1 represents a pairing between \( o_i \) and \( o_j \) and \(-1\) represents a non-pairing between \( o_i \) and \( o_j \)). Therefore, the proposed method for solving RL simultaneously possesses the properties of \( \forall n_{ij} \), \( n_{ij} \in N \cup \{0\} \).
Table 5
Summary of sampled students.

<table>
<thead>
<tr>
<th>Number of the Categorized Students</th>
<th>Department of Information Management (represented by O_0)</th>
<th>Department of Industrial Management (represented by O_1)</th>
<th>Department of Electrical Engineering (represented by O_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High scoring</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Average scoring</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Low scoring</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 6
Fuzzy extensions of RI between two different object data sets.

<table>
<thead>
<tr>
<th>(dataset, data set)</th>
<th>(O_0, O_2)</th>
<th>(O_0, O_1)</th>
<th>(O_0, O_1)</th>
<th>(O_0, O_2)</th>
<th>(O_0, O_2)</th>
<th>(O_0, O_1)</th>
<th>(O_0, O_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campello [8] (FRC for short)</td>
<td>0.68</td>
<td>0.7668</td>
<td>0.7135</td>
<td>0.7397</td>
<td>0.6804</td>
<td>0.8401</td>
<td></td>
</tr>
<tr>
<td>Brouwer [7] (FRB for short)</td>
<td>0.6820</td>
<td>0.7557</td>
<td>0.6880</td>
<td>0.7396</td>
<td>0.7144</td>
<td>0.8329</td>
<td></td>
</tr>
<tr>
<td>Anderson et al. [1] (FRA for short)</td>
<td>0.6038</td>
<td>0.6801</td>
<td>0.6334</td>
<td>0.6558</td>
<td>0.6093</td>
<td>0.7489</td>
<td></td>
</tr>
<tr>
<td>Hullermeier et al. [19] (FRH for short)</td>
<td>0.8421</td>
<td>0.8887</td>
<td>0.8048</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>FCR (our method)</td>
<td>0.7466</td>
<td>0.7134</td>
<td>0.6725</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

For example, a, b, c, d ∈ N U {0}, and ∀M, RI (M, M) = 1, for instance, FGR(E_1, E_1) = 1 and FGR(E_2, E_2) = 1.

We next explain why FGR(E_1, (3)p^{(1)}) = FGR(E_1, (4)p^{(1)}).

When we evaluate FGR(E_1, (3)p^{(1)}) and FGR(E_1, (4)p^{(1)}), their denominators are both the same and equal a + b + c + d = E^2 \times C_2^2 + 1000 (where τ = 10). Although a = 108 and d = 548 in FGR(E_1, (3)p^{(1)}) are different from a = 208 and d = 656 in FGR(E_1, (4)p^{(1)}), a + d = 656 in FGR(E_1, (3)p^{(1)}) is exactly the same as a + d = 656 in FGR(E_1, (4)p^{(1)}). Similarly, FGR(E_2, (1)p^{(1)}), FGR(E_2, (3)p^{(1)}), FGR(E_2, (4)p^{(1)}), FGR(E_2, (5)p^{(1)}), and FGR(E_2, (2)p^{(2)}) are all equal with the same reason. Of course, FGR(E_1, (1)p^{(1)}) = FGR(E_1, (2)p^{(2)}) and FGR(E_2, (1)p^{(1)}) = FGR(E_2, (1)p^{(2)}) because of (1)p^{(1)} = (1)p^{(2)}.

Example 2.

In this example, we use a real dataset to compare various fuzzy extensions of RI that had been reviewed in Section 2. Furthermore, we explain why the proposed method for the similarity of fuzzy partitions is more suitable than those existing fuzzy extensions of RI. Additionally, the purpose of providing this example is to consider whether students in different calculus courses taught by the same professor have similar learning strategies. Sternberg [30] described learning strategy as the method, activities, planning, and course of action used by a learner during the learning process. To educators, a deep understanding of the differences in learning strategies among different classes is necessary for facilitating teachers in improving teaching methods and creating effective learning environments. However, human behavior, thought, and language are subjective, complicated, diverse, and unclear. Therefore, both Law [25] and Ye [43] agreed that applying fuzzy theory is suitable for researching educational and psychological topics. In a Calculus course, a questionnaire regarding strategies for learning Calculus, which involved items regarding comprehension, practice problems, focus during class, and study plans (a list of the questionnaire items appears in Appendix A) that was translated from Chinese in [49] is modified to incorporate a membership function. A sample from first-year University students were asked to complete the questionnaire. The final data dimensions of each student oi were equal to 4 (i.e., o_i ∈ R^4). In an object o_i = (o_{i1}, o_{i2}, o_{i3}, o_{i4}), o_{i1} represents the comprehension measurement number of oi; o_{i2} represents the practice-problems measurement number of oi; o_{i3} represents the focus-during-class measurement number of oi; and o_{i4} represents the study-plan measurement number of oi. Because the questionnaire used a 5-point Likert scale, o_{ij} ∈ [1, 5] for all i and j = 1, 2, 3, 4. In each of three courses, a random sample of 10 students was selected. The students were categorized as high scoring, average scoring, and low scoring based on whether the students scored between 70 and 100, 50–70, or less than 50 on the mid-term and final examinations during the first semester of 2012. Table 5 presents a summary of the selected sample students.

Because we ask students to fill questionnaires with a form of membership values, we use a fuzzy partition to divide the object data sets. In general, Euclidean distance is the most frequently used for the data sets of higher dimension or spherical cluster data sets. Therefore, the fuzzy c-means (FCM) algorithm using Euclidean distance (see Bezdek [5] and Yang [38]) is used in this example. Because the students in each class are classified according to high, average, or low scores, regardless of the fuzzy partition used, the number of clusters k = 3. Let O_0 represent the object data set where the O_i is partitioned by FCM, t = A, B, C. The results are listed in Table 6.

From Table 6, we have the following results:

In FRC, FRB, FRA, we observe an unusual situation where the fuzzy similarity of the same data set (under the same clustering algorithm, i.e., FCM using Euclidean distance) is possibly lower than the fuzzy similarity index of different data sets (under the same clustering algorithm, i.e., FCM using Euclidean distance). For instance,

FRC(O_0, O_0) = 0.7397 < FRC(O_0, O_C) = 0.7668,

FRB(O_0, O_0) = 0.7396 < FRB(O_0, O_C) = 0.7557,

FRA(O_0, O_0) = 0.6558 < FRA(O_0, O_C) = 0.6801, etc.

Both methods developed by Hullermeier et al. [19] and the proposed method do not appear this abnormality because FRH(O_0, O_0) = 1 > FRH(O_0, O_C) = 0.8887 and FGR(O_0, O_0) = 1 > FGR(O_0, O_C) = 0.7134. But, the method of Hullermeier et al. [19] is still not suitable for fuzzy partition. For example, assume M_{SP}^{(1)} = \begin{bmatrix} 0.2 & 0.5 \\ 0.8 & 0.5 \end{bmatrix} and M_{SP}^{(2)} = \begin{bmatrix} 0.1 & 0.4 \\ 0.9 & 0.6 \end{bmatrix} are two fuzzy partitions of the data set O = \{o_1, o_2\}. Clearly, M_{SP}^{(1)} is not equal to M_{SP}^{(2)}. However, improper results such as FRH(M_{SP}^{(1)}, M_{SP}^{(2)}) = 1 occurred, nevertheless, FGR(M_{SP}^{(1)}, M_{SP}^{(2)}) = 0.8200 ≠ 1. In other words, although several researchers [17,8,19] had made fuzzy extensions of the Rand index, their methods may produce some unsuitable results in some statuses. In reality, the sign relation matrix used by the proposed FGR divides the object unit of the fuzzy partition matrix into smaller units. The fuzzy partition matrix for the processed smaller units is the crisp partition matrix. The fundamental essence is the same as RI. Therefore, our method
can avoid the loss of information and is a good choice for fuzzy partitions.

From FGRI, the calculus learning strategies of Information Management students are similar to those of Industrial Management students. This result is logical and reflects actual conditions. These two departments are within the School of Business (the Department of Electrical Engineering is in the School of Science), in which more female than male students are enrolled (the reverse is true in the School of Science). In addition, business students do not possess the same comprehension of mathematics as science students. Therefore, in studying the mathematics class calculus, business students differ slightly from electrical engineering students in comprehension and other learning strategies. Furthermore, the provided data show that the proposed method is successfully generalized to the treatment of similarity between two different object data sets with the same cardinal number and the same partition method. However, those results of other scholars do not match with the actual situation because of the above improper reasons. Therefore, the proposed FGRI provides a useful evaluation method that can be applied in future practical studies in various fields according to actual demands.

**Example 3.** Although FCM using Euclidean distance is effective when applied to spherical cluster data sets or the data sets of higher dimension, practical information is not always spherical data set. Gustafson and Kessel [14] proposed that Mahalanobis distance took place of Euclidean distance, called the GK algorithm. In practice, FCM using Mahalanobis distance is appropriate when applied to ellipse cluster data sets. However, none of the algorithms is perfectly valid when the data set contains spherical clusters and ellipse clusters. Therefore, combining FCM using Euclidean distance and FCM using Mahalanobis distance as a cluster ensemble is a useful clustering framework. To evaluate the similarity between two fuzzy cluster ensembles is important. In this example, we use a simple two-dimensional data set with clearly separated clusters to explain the similarity between two fuzzy cluster ensembles. And, the positive definite matrix of Mahalanobis distance is the inverse of covariance matrix. The data set $O$ as follows.

Assume $\theta(O) = 21$, $o_i \in R^2$ for all $i$ and $k = 4$. The object data set $O = \{O_1, O_2, O_3, \ldots, O_{21}\} = ((3, 5), (4, 7), (5, 4), (5, 5), (4, 5), (1, 14), (2, 12), (3, 13), (4, 11), (5, 12), (1, 15), (9, 6), (8, 8), (10, 7), (11, 3), (11, 5), (9, 14), (9, 15), (10, 14), (10, 16), (11, 15))$, where $o_1, \ldots, o_5 \in S_1$, $o_6, \ldots, o_{11} \in S_2$, $o_{12}, \ldots, o_{16} \in S_3$, $o_{17}, \ldots, o_{21} \in S_4$ as shown in Fig. 3.

**Fig. 3.** Data points of the object data set $O$.

From Fig. 3, we can clearly see that the scatters of $S_1$ and $S_4$ are spherical, but $S_2$ and $S_3$ are ellipse. Thus, we use the FCM of Euclidean distance and also use the FCM of Mahalanobis distance. We then consider the fuzzy cluster ensemble consisting of FCM of Euclidean distance and the FCM of Mahalanobis distance. Let $M_k^{(E)}$ indicate the fuzzy partition matrix of the fuzzy partition $P_k^{(E)}$ using the FCM, and let $M_k^{(M)}$ indicate the fuzzy partition matrix of fuzzy partition $P_k^{(M)}$ using the GK. Let $M_k^{(em)}$ indicate the membership matrix of the fuzzy cluster ensemble $E_k$ consists of $P_k^{(E)}$ and $P_k^{(M)}$, where $k$ is the number of clusters. Assume that $M_k$ indicates the crisp partition matrix of the crisp partition $P$, where the number of clusters is equal to 4. Then we can find FGRI by using $M_k^{(E)}$, $M_k^{(M)}$, $M_k^{(em)}$ and $M_k$. These FGRI values for different situations are shown in Tables 7–9.

From Tables 7–9, we can find the following two results:

(a) Regardless of different fuzzy clustering methods (containing fuzzy partitions $P_k^{(E)}$ and $P_k^{(M)}$), the object data set $O$ is successfully calculated to be closest to the crisp reference partition (i.e. $P$) when the number of clusters equals 4 (i.e. $k = 4$). This result corresponds with the expected results.

(b) The fuzzy cluster ensemble $E_k$ is closest to the crisp reference partition $P$, i.e. $FGRI(E_k, P) < FGRI(E_4, P)$, for all $k \neq 4$.

Furthermore, if $k = 4$ and the fuzzy cluster ensemble $E_4$ is considered as the fuzzy reference cluster ensemble, we have $FGRI(E_4, P_4^{(E)}) = 0.9814 < FGRI(E_4, P_4^{(M)}) = 0.9908$. This result tells us that $P_4^{(E)}$ is closer to the fuzzy reference cluster ensemble $E_4$ than $P_4^{(M)}$. It means the fuzzy cluster ensemble $E_4$ is more similar to the FCM with the Mahalanobis distance. Additionally, we can also find the result $FGRI(P_4^{(E)}, E_4^{(E)}) = 0.9725$ by using sign relation matrices. In other words, we can use the proposed fuzzy generalized rand index FGRI for several cases as shown in Fig. 4.

![Fig. 4. Different cases used by FGRI.](image-url)
5. Conclusions and discussion

The first contribution of this paper is the proposal of an evaluation method for cluster ensemble based on the proposed fuzzy generalized Rand index (FGRI). This proposal is good as an evaluation measure for cluster ensembles. The second contribution of this paper is that its method can avoid those irrationalities, such as ARImp for handling the ARI between a cluster ensemble and a crisp partition, and ARImm for the ARI between two cluster ensembles, proposed by Zhang et al. [44]. The third contribution of this paper is that the proposed method not only is suitable for handling the similarity of cluster ensembles, but also can be used to evaluate the similarity of fuzzy partitions, fuzzy cluster ensembles, and different subject data sets with the same cardinal number and the same partition method.

In cluster ensembles, the generated consensus matrices are intuitively used for evaluation measurement, such as the proposal in Zhang et al. [48]. However, it does not work for fuzzy cluster ensembles so that we propose the use of the membership matrices to substitute for the consensus matrices. Furthermore, to retain the relationship between an object and clusters (partitions) with the only two states: “belong to” or “not belong to”, i.e. yes or no, a new idea of a sign relation matrix is proposed. By using the trace and product of matrix, we successfully propose a suitable method FGRI to compute the similarity of nine cases: a cluster ensemble and a crisp partition, two cluster ensembles, a fuzzy partition and a crisp partition, two fuzzy partitions, a fuzzy cluster ensemble and a crisp partition, a fuzzy cluster ensemble and a cluster ensemble, a fuzzy cluster ensemble and a fuzzy partition, two fuzzy cluster ensembles, two different object data sets with the same cardinal number and the same partition method.

Although several researchers [1, 7, 8, 19] had proposed different fuzzy extensions of FIM, their methods are not suitable for fuzzy partitions. In this paper we make comparisons of the proposed method with those existing methods to demonstrate the efficiency of the proposed method. It is seen that the results from the proposed method are highly rational with strong potential for applications in practical scenarios. There are various cluster ensemble methods to be applied in gene expression data in the literature (see Refs. [20, 28, 45, 46]). In our further research, we will apply the proposed method to evaluate these cluster ensembles, especially for gene expression data.

Acknowledgements

The authors would like to thank the anonymous referees for their helpful comments in improving the presentation of this paper. This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant: MOST 103-2118-M-033-001-MY2.

Appendix A.

Calculation learning strategies questionnaire items and categories [49].

<table>
<thead>
<tr>
<th>Category</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehension</td>
<td>When I solve the problems, I can follow the steps taken by others, but I do not necessarily understand the reasoning.</td>
</tr>
<tr>
<td>Comprehension</td>
<td>I frequently find that after reading my calculus notes, I am still confused.</td>
</tr>
<tr>
<td>Comprehension</td>
<td>I understand my notes, but I still do not understand how to solve the problems.</td>
</tr>
<tr>
<td>Comprehension</td>
<td>I am frequently too busy taking notes during class to think about the content.</td>
</tr>
<tr>
<td>Comprehension</td>
<td>I frequently give up on calculus because I do not understand calculus.</td>
</tr>
</tbody>
</table>

References