Applied Soft Computing 57 (2017) 225–234

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

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

Chiou-Cherng Yeh[a,b], Miin-Shen Yang[a,*]

[a] Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan
[b] Teaching Center of Natural Science, Minghsin University of Science and Technology, Hsinchu, Taiwan

ARTICLE INFO

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

ABSTRACT

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 (*ARI*) 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.

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 a 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.

[*] Corresponding author.
E-mail address: msyang@math.cycu.edu.tw (M.-S. Yang).

http://dx.doi.org/10.1016/j.asoc.2017.03.030
1568-4946/© 2017 Elsevier B.V. All rights reserved.

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 measure 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$P^{(r)} = \left\{S_1^{(r)}, S_2^{(r)}, \cdots, S_{k_r}^{(r)}\right\}$ where $\overset{k_r}{\underset{h=1}{\cup}} S_h^{(r)} = O$and$S_h^{(r)} \cap S_{h'}^{(r)} = \varphi$ 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 $n_{u\cdot} = \sum_{v=1}^{k_2} n_{uv}$, $n_{\cdot v} = \sum_{u=1}^{k_1} n_{uv}$, 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:

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

$$= \frac{\binom{n}{2} + 2\sum_{u=1}^{k_1}\sum_{v=1}^{k_2}\binom{n_{uv}}{2} - \left[\sum_{u=1}^{k_1}\binom{n_{u.}}{2} + \sum_{v=1}^{k_2}\binom{n_{.v}}{2}\right]}{\binom{n}{2}} \tag{1}$$

Table 1
The contingency table N.

Cluster	$S_1^{(2)}$	$S_2^{(2)}$	$\cdots$	$S_4^{(2)}$	$\sum$
$S_1^{(1)}$	n_{11}	n_{12}	$\cdots$	n_{1k_2}	$n_{1.}$
$S_2^{(1)}$	n_{21}	n_{22}	$\cdots$	n_{2k_2}	$n_{2.}$
$\vdots$	$\vdots$		$\cdots$		$\vdots$
$S_{k_1}^{(1)}$	n_{k_11}	n_{k_12}	$\cdots$	$n_{k_1k_2}$	$n_{k_1.}$
$\sum$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.k_2}$	n

Assume that the crisp partition matrix $M_C^{(r)} = \left[m_{hi}^{(r)}\right]_{k_r \times n}$ of $P^{(r)}, r = 1, 2$is defined as $m_{hi}^{(r)} = \begin{cases} 1 & \text{if} o_i \in S_h^{(r)} \\ 0 & \text{if} o_i \notin S_h^{(r)} \end{cases}$. Let $N = (M_C^{(1)})(M_C^{(2)})^T$, 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

$$ARI(P^{(1)}, P^{(2)}) = \frac{RI - E(RI)}{\max(RI)\text{-}E(RI)}$$

$$= \frac{\sum_{u=1}^{k_1}\sum_{v=1}^{k_2}\binom{n_{uv}}{2} - \dfrac{2\sum_{u=1}^{k_1}\binom{n_{u.}}{2}\sum_{v=1}^{k_2}\binom{n_{.v}}{2}}{n(n-1)}}{2\left[\sum_{u=1}^{k_1}\binom{n_{u.}}{2} + \sum_{v=1}^{k_2}\binom{n_{.v}}{2}\right] - \dfrac{2\sum_{u=1}^{k_1}\binom{n_{u.}}{2}\sum_{v=1}^{k_2}\binom{n_{.v}}{2}}{n(n-1)}} \tag{2}$$

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

$$l_0 = \sum_{u=1}^{k_1}\sum_{v=1}^{k_2}\binom{n_{uv}}{2},\ l_1 = \sum_{u=1}^{k_1}\binom{n_{u.}}{2},\ l_2 = \sum_{v=1}^{k_2}\binom{n_{.v}}{2},\ l_3 = \frac{2 \cdot l_1 \cdot l_2}{n(n-1)}.\text{ Then}$$

$$ARI(P^{(1)}, P^{(2)}) = \frac{l_0 - l_3}{\frac{1}{2}(l_1 + l_2)\text{-}l_3} = \frac{a - \frac{2(b+a)(c+a)}{n(n-1)}}{\frac{1}{2}(b + c - 2a) - \frac{2(b+a)(c+a)}{n(n-1)}} \tag{3}$$

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 $P = \left\{S_1, S_2, \cdots, S_k\right\}$, where $\overset{k}{\underset{h=1}{\cup}} S_h = OandS_h \cap S_{h'} = \varphi$ for all $h \neq h'$. Then the relation matrix $R = \left[r_{ij}\right]_{n \times n}$of O for P is defined as follows:

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

Let the co-association matrix A for P be $A = R - I = \left[a_{ij}\right]_{n \times n}$,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 ${}^{(1)}P, {}^{(2)}P, \cdots$, and ${}^{(q')}P$. We know that a cluster ensemble is used to combine these q' crisp partitions ${}^{(1)}P, {}^{(2)}P, \cdots$, and ${}^{(q')}P$ into a useful clustering framework. Let the co-association matrix for ${}^{(w)}P$ be defined as ${}^{(w)}A = \left[{}^{(w)}a_{ij}\right]_{n \times n}$where $w = 1, \cdots, q'$. A consensus matrix of cluster ensemble can be constructed as