IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 13, 2017, accepted August 7, 2017, date of publication August 17, 2017, date of current version September 19, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2741221

Generalized Pair-Counting Similarity Measures for Clustering and Cluster Ensembles

SHAOHONG ZHANG[1], ZONGBAO YANG[1], XIAOFEI XING[1], (Member, IEEE), YING GAO[1], DONGQING XIE[1], AND HAU-SAN WONG[2]

[1]Department of Computer Science, Guangzhou University, Guangzhou 510006, China
[2]Department of Computer Science, City University of Hong Kong, Hong Kong

Corresponding author: Shaohong Zhang (zimzsh@qq.com)

This work was supported in part by the Scientific and Technological Project of Guangzhou, China, under Project 201607010053, Project 201607010191, and Project 201707010284, in part by the Outstanding Young Teachers in the Higher Education Institutions of Guangdong Province, China, under Project Yq201401, in part by the Teaching Reform Project of Guangzhou University under Project JY2016033, in part by the 2016 Education Reform Project of the Guangdong Province, through the Project Research on Construction and Mining methods in Knowledge Graphs of Computer Sciences Courses, under Project 426, File [2016]236, in part by the Natural Science Foundation of Guangdong, China, under Grant 2014A030313524 and Grant 2016A030313540, and in part by the Science and Technology Projects of Guangdong Province, China, under Project 2016B010127001, and in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project CityU 11300715, and in part by the City University of Hong Kong under Project 7004674.

ABSTRACT In this paper, a number of pair-counting similarity measures associated with a general formulation of cluster ensemble are proposed. These measures are formulated based on our motivation to evaluate the consistency between an individual clustering solution and a cluster ensemble solution, or that between different cluster ensemble solutions, in a uniform manner. A number of criteria are proposed for the comparison of these generalized measures, from both the perspectives of theoretical analysis and experimental validation. We identify their different behaviors and their correlations in different scenarios of traditional clustering solutions and cluster ensembles, with the hope that the results of these studies could 1) serve as important criteria for the design and selection of evaluation measures for clustering solutions, and 2) provide explanations for ambiguous clustering results in related scenarios. Experiments with both synthetic and real data sets are conducted to verify our findings.

INDEX TERMS Clustering evaluation, cluster ensembles, similarity measures.

I. INTRODUCTION

With the development of modern data acquisition techniques, the problem of extracting useful pattern information from those data becomes more important. Examples include microarray data, web site visit log data, and data resulting from social network analysis. Clustering is one of the most important approaches for organizing such data in many practical applications, and different clustering algorithms have been proposed for general or specific tasks. However, due to different characteristic properties of these data sets, such as different data distributions, different data sizes, different amount of noise, and different data preprocessing methods adopted, a single clustering method cannot guarantee to behave well in all scenarios. Motivated by the successful development of ensemble approaches in supervised classification, such as Adaptive Boosting (AdaBoost) [1] or Random Forest [2], ensemble techniques in unsupervised clustering applications have attracted great interests in machine learning communities [3]–[12]. Specifically, the unsupervised ensemble framework, which is usually referred to as cluster ensemble [3], creates a consensus solution from multiple clustering solutions and usually achieves more accurate results. In general, the individual clustering solutions can be generated from different perspectives, such as different data sample subsets [5], [6], different data feature subsets [4], [5], or different clustering algorithms [3], [13]. Once the individual partitions are created, the consensus clustering solution can be derived using co-association based methods [3], [5], [7] or graph partitioning based methods [3], [4]. For the first category of cluster ensemble methods, a co-association matrix is generated from each individual partition, where the (i, j)-th entry of the matrix denotes whether the i-th data point and the j-th data point belong to the same cluster in the partition. After that, all the co-association matrices are summed to form a

2169-3536 © 2017 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

consensus matrix. This matrix can thus serve as a pairwise similarity matrix, and a consensus solution can be derived using suitable clustering algorithms [5], [7] or graph cut algorithms [3]. The second category of cluster ensemble methods generates a graph representation based on the relationship between samples/clusters/partitions from the individual clustering solutions, and searches for a final consensus clustering solution using a suitable graph partitioning algorithm [4], [5]. In general, cluster ensemble methods generate more stable and accurate solutions compared to the individual clustering solutions.

In many cases the accurate evaluation of the cluster ensemble results requires measures associated with ground truth information, commonly known as external measures. Popular external clustering evaluation measures can roughly be divided into three categories:(i) Pair-counting measures, which are calculated based on the extent to which the cluster and class memberships of pairs of data points agree or disagree. Well-known measures in this category include Rand Index [14], Adjusted Rand Index [15], Fowlkes-Mallows Index [16] and Jaccard Index [17]. More comprehensive studies on these pair-counting measures can be found in [18]–[21]; (ii) Information-theoretic measures, which are based on various forms of entropic measures from information theory. Well-known measures in this category include mutual information [22], purity [23], Normalized Mutual Information [3], and Variation of Information [24]. More recent studies on the measures in this category can be found in [25]–[27]; and (iii) Set-matching measures, which characterize the extent of similarity between clusters in two partitions based on set theoretic measures. Well-known measures in this category include the van Dongen criterion [28], the $\mathcal{H}$ criterion [29], and the $\mathcal{L}$ criterion [30].

However, in many practical scenarios, the ground truth labels are not available [31], or in other forms rather than the label information. Representative examples include pairwise linkage in web data [32], [33], structural relationship in social network [34], and semantic similarity in gene regulation network and Gene Ontology [35], [36]. In addition, to evaluate the improvement of quality resulting from the formation of the final cluster ensemble solution compared with the individual partitions, and the diversity among the individual partitions, the average similarity between pairs of individual partitions based on various external measures are generally used, such as Average Normalized Mutual Information (ANMI) [3], [5] and Pairwise Normalized Mutual Information (PNMI) [5], [37]. Evaluation based on external measures will become less applicable when these individual partitions are represented in a fuzzy form in terms of membership functions (instead of a crisp partition), or when the individual partitions consist of different subsets of the data samples, or when the partitions are less accessible due to privacy and/or security concerns. Therefore, it is important to develop effective measures to evaluate the consistency between partitions and pairwise similarity matrices, and that between different pairwise similarity matrices for both individual clustering solutions and cluster ensembles, in a similar spirit to the adoption of external measures above. Motivated by the Adjusted Rand Index (ARI) and its fuzzy extension [19], we have proposed two new generalized ARI measures in our recent works [20], [21]. However, there are few previous studies about the generalization of other measures. It is interesting to note that other measures in the pair-counting category can also be generalized, while generalization of the measures in the other two categories is difficult to proceed. Although there are a lot of related works on clustering measures under the traditional clustering scenarios [24], [31], [38]–[40], there are, to the best of our knowledge, no previous studies on the behaviors of these measures in the more general cluster ensemble scenario. Also, there is a lack of comparison of these measures in both scenarios. More importantly, there are no studies which provide comprehensive investigation of these generalized measures in scenarios of clustering and of cluster ensembles. The work in this paper thus attempts to bridge these gaps, with the hope that the results of these studies could (i) serve as important criteria for the design and selection of evaluation measures for different scenarios, and (ii) provide explanations for ambiguous clustering results in related scenarios.

A. OUR WORK AND CONTRIBUTION IN THIS PAPER

In this paper, we generalize 21 pair-counting measures to evaluate the consistency between individual partitions and pairwise similarity matrices, or between two pairwise similarity matrices, in a unified framework. To compare these generalized measures in the scenarios of traditional clustering solutions and cluster ensembles, we propose a number of criteria and investigate their properties, from both the perspectives of theoretical and experimental analysis. To our best knowledge, there are no previous studies which provide comprehensive investigation of these generalized measures in scenarios of clustering and of cluster ensembles using these sets of proposed properties. In addition, different applications of these generalized measures in practical cluster ensemble scenarios are also discussed.

B. ORGANIZATION OF THIS PAPER

The rest of the paper is organized as follows. Section 2 provides a brief introduction of cluster ensemble and related pair-counting measures. Section 3 presents a unified framework for generalizing these pair-counting measures. Section 4 proposes a number of effective criteria to compare various generalized pair-counting measures based on theoretical analysis. Section 5 describes experimental results to verify the properties of the generalized measures and to demonstrate their effectiveness in practical cluster ensemble tasks. Section 6 concludes the paper.

II. CLUSTER ENSEMBLES AND CLUSTERING EVALUATION MEASURES

We briefly introduce related background information about cluster ensemble based on the consensus matrix,

and introduce a number of pair-counting similarity measures for clustering solution evaluation. In this paper the two terms "clustering solution" and "partition" are used interchangeably.

A. CLUSTER ENSEMBLE BASED ON CONSENSUS MATRIX

We now introduce cluster ensemble based on the consensus matrix. Given a data set $X = \{x_i\}_{i=1}^{N}$ with N data points, a partition P divides X into a set of mutually disjoint clusters $\{P_k\}_{k=1}^{K}$. For the partition P, an $N \times N$ co-association matrix can be constructed as follows:

$$M_{ij} = \begin{cases} 1 & \text{if } \exists\, k,\ x_i \in P_k \text{ and } x_j \in P_k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Note that the diagonal entries of M will always be 1 because each point is itself a member of the cluster to which it belongs.

Given a set of partitions $\{P^{(l)}\}_{l=1}^{L}$, the $N \times N$ consensus matrix can be computed as the average of their corresponding co-association matrices using

$$\mathcal{M} = \frac{1}{L}\sum_{l=1}^{L} M^{(l)} \tag{2}$$

Note that this consensus matrix can be regarded as a fuzzy generalization of the individual co-association matrices.

B. PAIR-COUNTING SIMILARITY MEASURES

A number of cluster evaluation measures have been proposed in various pair-counting similarity forms. Specifically, let $P = \{P_1, P_2, \cdots, P_{K^{(P)}}\}$ and $Q = \{Q_1, Q_2, \cdots, Q_{K^{(Q)}}\}$ be two partitions on a data set $X = \{x_i\}_{i=1}^{N}$ with N entities, $P(x_i)$ and $Q(x_i)$ be the corresponding class labels for the data point x_i in partition P and Q respectively, and $\delta()$ be the indicator function with $\delta(true) = 1$ and $\delta(false) = 0$, four different factors are defined as follows

$$\begin{aligned} a &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \delta(P(x_i) = P(x_j)\ and\ Q(x_i) = Q(x_j)) \\ b &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \delta(P(x_i) = P(x_j)\ and\ Q(x_i) \neq Q(x_j)) \\ c &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \delta(P(x_i) \neq P(x_j)\ and\ Q(x_i) = Q(x_j)) \\ d &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \delta(P(x_i) \neq P(x_j)\ and\ Q(x_i) \neq Q(x_j)) \end{aligned} \tag{3}$$

Note that the sum λ of these four factors equals the number of all possible point pairs

$$\lambda = a + b + c + d = \binom{N}{2} = \frac{N(N-1)}{2} \tag{4}$$

Most of the proposed cluster evaluation measures in various pair-counting similarity forms are based on these five factors, and they are bounded in the intervals $[0, 1]$ or $[-1, 1]$. In this paper, we shall focus on such bounded evaluation measures as shown in Table 1, where details of their names, notations, references, original definitions and ranges are included.

TABLE 1. Pair-counting similarity measures investigated in this paper.

Name	Notation	Reference	Definition	Range
Adjusted Rand Index	ARI	[15]	$\frac{a-\frac{(a+b)(a+c)}{\lambda}}{0.5(a+b+a+c)-\frac{(a+b)(a+c)}{\lambda}}$	(-1, 1]
Baulieu	B	[41]	$\frac{\lambda^2-\lambda(b+c)+(b-c)^2}{\lambda^2}$	[0, 1]
Czekanowski	CZ	[42]	$\frac{2a}{2a+b+c}$	[0, 1]
Fowlkes-Mallows	FM	[16]	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[0, 1]
Gamma (Γ)	G	[15]	$\frac{\lambda a-(a+b)(a+c)}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	[-1, 1]
Goodman and Kruskal	GK	[43]	$\frac{ad-bc}{ad+bc}$	[-1, 1]
Gower and Legendre	GL	[44]	$\frac{a+d}{a+0.5(b+c)+d}$	[0, 1]
Hamann	H	[45]	$\frac{(a+d)-(b+c)}{\lambda}$	[-1, 1]
Jaccard	J	[17]	$\frac{a}{a+b+c}$	[0, 1]
Kulczynski	K	[46]	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$	[0, 1]
McConnaughey	MC	[47]	$\frac{a^2-bc}{(a+b)(a+c)}$	[-1, 1]
Pearson	P	[18]	$\frac{ad-bc}{(a+b)(a+c)(c+d)(b+d)}$	[-1, 1]
Peirce	PE	[48]	$\frac{ad-bc}{(a+c)(b+d)}$	[-1, 1]
RAND	R	[14]	$\frac{a+d}{\lambda}$	(0, 1]
Russel and Rao	RR	[49]	$\frac{a}{\lambda}$	[0, 1]
Rogers and Tanimoto	RT	[50]	$\frac{a+d}{a+2(b+c)+d}$	[0, 1]
Sokal and Sneath 1	SS1	[51]	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$	[0, 1]
Sokal and Sneath 2	SS2	[51]	$\frac{a}{a+2(b+c)}$	[0, 1]
Sokal and Sneath 3	SS3	[51]	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	[0, 1]
Wallance 1	W1	[16]	$\frac{a}{a+b}$	[0, 1]
Wallance 2	W2	[16]	$\frac{a}{a+c}$	[0, 1]

III. GENERALIZED PAIR-COUNTING SIMILARITY MEASURES FOR CLUSTER ENSEMBLE: A UNIFIED FRAMEWORK

Generalized pair-counting similarity measures for cluster ensemble can be derived from the original definitions based on the corresponding consensus matrices (or co-association matrices for partitions as a special case) following the methods adopted in [19]–[21]. Specifically, given two consensus matrices $\mathcal{M}^{(P)}$ and $\mathcal{M}^{(Q)}$ for two cluster ensembles, the factors defined in Eq. (3) can be computed as follows

$$\begin{aligned} a &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \mathcal{M}_{ij}^{(P)}\mathcal{M}_{ij}^{(Q)} \\ b &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \mathcal{M}_{ij}^{(P)}(1-\mathcal{M}_{ij}^{(Q)}) \\ c &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} (1-\mathcal{M}_{ij}^{(P)})\mathcal{M}_{ij}^{(Q)} \\ d &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} (1-\mathcal{M}_{ij}^{(P)})(1-\mathcal{M}_{ij}^{(Q)}) \end{aligned} \tag{5}$$

Based on the new factors in Eq. (5), generalized pair-counting similarity measures can be derived from the same definitions in Table 1, with the original measures obtained from a number of references [14]–[16], [16]–[18], [41]–[51].

It is interesting that the factors computed in Eq. (5) are identical to those defined in Eq. (3) when the two consensus matrices reduce to co-association matrices (i.e., computed from two partitions). In this case, the original pair-counting similarity measures for two partitions can be viewed as special cases of the newly generalized ones.

IV. COMPARISON CRITERIA FOR GENERALIZED PAIR-COUNTING SIMILARITY MEASURES

Given the various measures proposed above, a further study of their properties is important. To our best knowledge, there are no studies on these generalized measures in a wide context from traditional clustering solutions to cluster ensembles in a uniform manner. It can be expected that these measures might possess a set of core properties, and at the same time have their own special properties. In view of this, we perform a theoretical analysis of these generalized measures and their properties in this section. Related experimental study will further be conducted on both simulated and real data sets in the experiment section.

Given two consensus matrices $\mathcal{M}^{(P)}$ and $\mathcal{M}^{(Q)}$ for two cluster ensembles (also including the special case of clustering partitions) and the measure $sim(\mathcal{M}^{(P)}, \mathcal{M}^{(Q)})$, we propose to study a number of different properties including:

- **P1: Symmetry.** Whether $sim(\mathcal{M}^{(P)}, \mathcal{M}^{(Q)}) = sim(\mathcal{M}^{(Q)}, \mathcal{M}^{(P)})$?
- **P2: Possibility of distributed computation.** Whether these general measures can be computed in a distributed manner.
- **P3: Detection of Uncorrelatedness (DoU).** Whether the measures can converge to a baseline value in the case of two random partitions or two random ensembles.
- **P4: Measure of complementarity.** The complement of a partition (or an ensemble) is represented in the form of a matrix, where the summation of each non-diagonal entry of the matrix and that of the co-association matrix of the partition (of the consensus matrix of the ensemble) is 1. We would like to determine the relationship between a partition (or an ensemble) and its complement.
- **P5: Measure of self-similarity.** Whether the measures can readily evaluate the similarity between a partition or a cluster ensemble with itself.

Since most of the generalized similarity measures are derived from widely adopted clustering evaluation measures, it is difficult to explicitly compare their effectiveness in different scenarios. To our best knowledge, there are no previous studies on the comparison of generalized similarity measures in these multiple scenarios. Although it is difficult to perform a complete comparison from all perspectives, the aforementioned properties inherit the spirit of recent studies in cluster quality measures [18], [19], [25], [52] and cluster ensembles [3], [5], [37], [53], with the added advantage of providing a wider coverage to include both the cases of clustering and cluster ensemble:

- Property 1 (P1: Symmetry) requires that the measures should output the same similarity value for two cluster ensembles regardless of their ordering;
- Property 2 (P2: Possibility of distributed computation) explores the applicability of distributed computation for these measures, since distributed computation is one of the possible benefits of cluster ensembles [3];
- Property 3 (P3: Detection of Uncorrelatedness) investigates the behavior of these measures under both random clusterings or random ensembles, which has been mainly discussed in the case of clustering comparison [15], [18], [25] and briefly investigated for the generalized Adjusted Rand Index in cluster ensembles in our earlier work [20], [21];
- Property 4 (P4: Measure of complementarity) generalizes two goodness criteria (Homogeneity and Completeness) for clustering comparison in [52] to characterize complementary clustering solutions;
- Property 5 (P5: Measure of self-similarity) evaluates the capability of the generalized measures to compute the self-similarity of a cluster ensemble. This property can be used to study the diversity of cluster ensembles in a more general manner, which is widely believed to be important for the quality of cluster ensembles [5], [21], [37], [53].

These properties are discussed in detail in the following subsections. Finally, the overall comparison results are summarized and discussed with respect to previous findings at the end of this section.

A. PROPERTY 1 (P1): SYMMETRY

The symmetry property can be directly identified from the definitions of the measures. Referring to the measure definitions in Table 1, only some measures, such as Peirce (PE), Wallance 1 (W1) and Wallance 2 (W2), do not satisfy this condition. This also suggests that most researchers have this property in mind when designing these measures.

B. PROPERTY 2 (P2):POSSIBILITY OF DISTRIBUTED COMPUTATION

As discussed in one of our recent papers [21], we can effectively compute the generalized Adjusted Rand Index in a distributed manner. It is interesting to see that all the other generalized pair-counting similarity measures have this desirable property. Specifically, two consensus matrices can be split into several sub-matrices, and the factors associated with each sub-matrix can be computed in a distributed manner. This distributed computation approach for generalized pair-counting similarity measures is useful when the memory requirement is too large for one single machine, and/or data security requirement is important.

C. PROPERTY 3 (P3): DETECTION OF UNCORRELATEDNESS (DoU)

One of the important properties of a clustering evaluation measure that attracts great interest is how it compares two independent random partitions [15], [18], [20], [21], [25]. In general, random partitions are constructed by assigning data points to their clusters with a uniform distribution. Specifically, the number of clusters K is randomly selected from 2 to a maximum number K_{max} using a uniform distribution, and the assignment probabilities of a data point to any cluster are equal (i.e., $1/K$). In this case, desirable measure values should be close to a baseline value, for example zero. Among measures that have this property, the Adjusted Rand Index is the most notable one [15], [18], [20], [21], [25]. In our earlier study [21], we have proposed a generalized Adjusted Rand index between two independent random ensembles, each of which corresponds to a set of random partitions. Here we focus on the corresponding property of the other generalized measures in the context of comparing random partitions and random ensembles, and uniformly refer to it as ''Detection of Uncorrelatedness (DoU)''. From our study, we find that it is more accurate to view the baseline value as the mid-point of the measure value range for the different similarity measures rather than zero. Six measures, including ARI, GAMMA, GK, P, PE, and SS1, are found to have this desirable property. Among these six measures, the first five have a value range of $[-1, 1]$, with a baseline value of 0, while the range of SS1 is $[0, 1]$ with a baseline value of 0.5. We also provide a proof of this property for these six measures when applied to two independent random partitions and to two independent random ensembles.

1) PROPERTY 3a (P3a): DETECTION OF UNCORRELATEDNESS (DoU) BETWEEN TWO RANDOM PARTITIONS

ARI, GAMMA, GK, P, PE, and SS1 have the common property of DoU between two random partitions.

From the definitions of the measures, we can observe that for the first five measures, i.e., ARI, GAMMA, GK, PE, and P, they are in the form of $\frac{ad-bc}{pos}$, where pos is positive. It can be seen that

$$\begin{aligned}
&ad - bc \\
&= a\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(1 - M_{ij}^{(P)})(1 - M_{ij}^{(Q)}) \\
&\quad - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(P)}(1 - M_{ij}^{(Q)}) \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(1 - M_{ij}^{(P)})M_{ij}^{(Q)} \\
&= a\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(1 - M_{ij}^{(P)} - M_{ij}^{(Q)} + M_{ij}^{(P)}M_{ij}^{(Q)}) \\
&\quad -(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(P)} - a)(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(Q)} - a) \\
&= a\lambda - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(P)} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(Q)} \\
&= \lambda\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(P)}M_{ij}^{(Q)} - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(P)} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(Q)}
\end{aligned} \tag{6}$$

Given a particular random partition generated with the uniform distribution, say P with $K^{(P)}$ clusters, the probability that an entry in its co-association matrix $M^{(P)}$ equals 1 (i.e., for the corresponding pair of points to belong to the same cluster) can be determined as follows

$$\begin{aligned}
p^{(P)} &= p(M_{ij}^{(P)} = 1) = \frac{\binom{K^{(P)}}{1}}{K^{(P)}K^{(P)}} = \frac{1}{K^{(P)}}, \\
p(M_{ij}^{(P)} = 0) &= 1 - \frac{1}{K^{(P)}}
\end{aligned} \tag{7}$$

i.e., $p(M_{ij}^{(P)})$ is a Bernoulli distribution. In the case of two independent random partitions as discussed here, if N, the number of data points, is sufficiently large, we can obtain

$$\begin{aligned}
a &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(P)}M_{ij}^{(Q)} \approx \frac{N(N-1)}{2}E(M_{ij}^{(P)}M_{ij}^{(Q)}) \\
&= \lambda E[M_{ij}^{(P)}]E[M_{ij}^{(Q)}] = \lambda p^{(P)}p^{(Q)}
\end{aligned} \tag{8}$$

where $p^{(P)}$ and $p^{(Q)}$ are the probabilities for the entries in the co-association matrices $M^{(P)}$ and $M^{(Q)}$ to be 1, for the two random partitions P and Q respectively. The second step follows from the law of large numbers, and the third step is based on the uncorrelatedness of the two partitions P and Q.

Also

$$\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(P)} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(Q)} \approx \lambda E(M_{ij}^{(P)})\lambda E(M_{ij}^{(Q)}) = \lambda^2 p^{(P)} p^{(Q)} \quad (9)$$

Thus, the factor $(ad - bc)$ in Eq. (6) can be further simplified as follows:

$$\begin{aligned} &ad - bc \\ &= \lambda \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(P)} M_{ij}^{(Q)} - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(P)} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} M_{ij}^{(Q)} \\ &\approx \lambda(\lambda p^{(P)} p^{(Q)}) - \lambda^2 p^{(P)} p^{(Q)} = 0 \end{aligned} \quad (10)$$

As a result, for ARI, GAMMA, GK, PE, and P, their values are close to zero in the case of two random partitions, which is the mid-point of their ranges $[-1, 1]$.

On the other hand, for the case of $ad - bc = 0$, the value of SS1 is determined as follows:

$$\begin{aligned} &SS1(M^{(Q)}, M^{(P)})) \\ &= 0.25(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c}) \\ &= 0.25((\frac{a}{a+b} + \frac{d}{d+c}) + (\frac{a}{a+c} + \frac{d}{d+b})) \\ &= 0.25(\frac{2ad + ac + bd}{ad + bd + bc + ac} + \frac{2ad + ab + cd}{ad + cd + ab + bc}) \\ &= 0.25(\frac{2ad + ac + bd}{ad + bd + ad + ac} + \frac{2ad + ab + cd}{ad + cd + ab + ad}) \\ &= 0.5 \end{aligned} \quad (11)$$

The fifth step uses the substitution $bc = ad$. This deduction shows that the value of SS1 is 0.5 in the case of two random partitions, which is the mid-point of its range [0, 1].

2) PROPERTY 3b (P3b): DETECTION OF UNCORRELATEDNESS (DoU) BETWEEN TWO RANDOM ENSEMBLES

ARI, GAMMA, GK, P, PE, and SS1 have the common property of DoU between two random ensembles.

Similar to the proof above, we can also focus on the factor $ad - bc$. Note that in the case of random ensembles, the entries of the consensus matrices are the average of those of the co-association matrices associated with the individual random partitions (ref to Eq. (2)). In addition, we also know that the distribution of the co-association matrix entries is Bernoulli. Thus, we have

$$\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \mathcal{M}_{ij} \approx \lambda E[\mathcal{M}_{ij}] = \lambda \frac{1}{L}\sum_{l=1}^{L} p^{(l)} \quad (12)$$

$$\begin{aligned} a &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \mathcal{M}_{ij}^{(P)} \mathcal{M}_{ij}^{(Q)} \approx \frac{N(N-1)}{2} E(\mathcal{M}_{ij}^{(P)} \mathcal{M}_{ij}^{(Q)}) \\ &= \lambda E[\mathcal{M}_{ij}^{(P)}] E[\mathcal{M}_{ij}^{(Q)}] = \lambda(\frac{1}{L^{(P)}}\sum_{l_1=1}^{L^{(P)}} p^{(l_1)})(\frac{1}{L^{(Q)}}\sum_{l_2=1}^{L^{(Q)}} p^{(l_2)}) \end{aligned} \quad (13)$$

where the second steps of Eq. (12) and Eq. (13) follow from the law of large numbers, and the third step of Eq. (13) results from the uncorrelatedness of the two ensembles. Similarly, we can derive the other terms as follows:

$$\begin{aligned} &ad - bc \\ &= \lambda \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \mathcal{M}_{ij}^{(P)} \mathcal{M}_{ij}^{(Q)} - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \mathcal{M}_{ij}^{(P)} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \mathcal{M}_{ij}^{(Q)} \\ &\approx \lambda^2(\frac{1}{L^{(P)}}\sum_{l_1=1}^{L^{(P)}} p^{(l_1)})(\frac{1}{L^{(Q)}}\sum_{l_2=1}^{L^{(Q)}} p^{(l_2)}) \\ &\quad - \lambda^2(\frac{1}{L^{(P)}}\sum_{l_1=1}^{L^{(P)}} p^{(l_1)})(\frac{1}{L^{(Q)}}\sum_{l_2=1}^{L^{(Q)}} p^{(l_2)}) \\ &= 0 \end{aligned} \quad (14)$$

Therefore, we can see that for two random ensembles, the measures ARI, GAMMA, GK, P, PE are close to zero, which is the middle of their bounds $[-1, 1]$, while the measure SS1 is close to 0.5.

Note that the other 15 measures do not output meaningfully converged values in the cases of two random partitions or two random ensembles, which will be shown in the experiment section.

D. PROPERTY 4 (P4): MEASURE OF COMPLEMENTARITY

It is interesting to note that cases where the generalized pair-counting similarity measures are below their baseline values were seldom discussed in previous works. Only very few examples were mentioned in the literature. For example, negative values of the Adjusted Rand Index (ARI) were discussed in [8] and [24]. Some negative value examples of ARI can also be found in [21] and [25]. To our best knowledge, no previous extensive studies have been conducted to investigate cases in which the similarity values are below their baseline. In this subsection we address these issues, propose related criteria, and discuss preliminary observations. In particular, we introduce definitions for the complement of a partition and a cluster ensemble, and identify well-behaved measures with values below their baseline when comparing partitions and their complements (or ensembles and their complements). We also provide further analysis to support our findings in the following subsections.

1) DEFINITION: COMPLEMENT OF A PARTITION AND A CLUSTER ENSEMBLE

It is difficult to define the corresponding complements directly for partitions or ensembles. Instead, we define them based on the co-association matrices for partitions, and consensus matrices for ensembles. Specifically, given an $N \times N$ co-association matrix M for a partition P with N points, or an $N \times N$ consensus matrix $\mathcal{M}$ for a cluster ensemble $\mathcal{E}$, their complements are defined as follows:

$$\begin{aligned} cpm(P) &= cpm(M) = 1_{N\times N} - M + E \\ cpm(\mathcal{E}) &= cpm(\mathcal{M}) = E - \mathcal{M} + 1_{N\times N} \end{aligned} \quad (15)$$

where E is the identity matrix and $1_{N\times N}$ is an $N \times N$ matrix with all entries being 1. (Similar definitions apply for $0_{N\times N}$ and $0.5_{N\times N}$ which appears later in this section).

2) PROPERTY 4a (P4a): THE MOST DISCRIMINATIVE COMPLEMENTARY PAIR

Given a data set, which complementary pair will be the most discriminative for a partition or an ensemble? It is not difficult to imagine the following special cases from a pair-counting perspective: (i) the fully-connected partition $P^{(F)}$ whose co-association matrix is $1_{N\times N}$, i.e., there is only one cluster in the partition and all the data points in this partition belong to the same cluster; and (ii) the singleton partition $P^{(S)}$ whose co-association matrix is the identity matrix, i.e., there are N data points as well as N clusters in $P^{(S)}$, and each data point belongs to a single cluster. An interesting related problem was discussed in [52] where two criteria were proposed to measure the goodness of a clustering measure between two partitions P and Q:

(i) **Homogeneity**, data assigned to a single cluster in partition Q should come from only a single cluster in partition P; (ii)**Completeness**, data in a single cluster of partition P should be assigned to a single cluster in partition Q.

It is interesting to observe that compared with $P^{(F)}$, any partition other than $P^{(F)}$ will satisfy the homogeneity criterion but violate the completeness criterion, while compared with $P^{(S)}$, any partition other than $P^{(S)}$ will satisfy the completeness criterion but violate the homogeneity criterion. Thus, a desirable measure should attain the minimum possible value when used to evaluate the similarity between these two different complementary pairs.

Since we know that some of the generalized measures are asymmetrical from subsection IV-A, we discuss this problem for two cases: (1) **Property 4a1 (P4a1)** $sim(P^{(F)}, P^{(S)})$, and (2) **Property 4a2 (P4a2)** $sim(P^{(S)}, P^{(F)})$. It is easy to identify those generalized measures which have this property by using the co-association matrices of $P^{(F)}$ and $P^{(S)}$, i.e., $1_{N\times N}$ and E, respectively. From Eq. (3), we can obtain that for the case of **P4a1**, $a = 0, b = \frac{N(N-1)}{2} = \lambda, c = 0, d = 0$. For the symmetric case (**P4a2**), i.e., when evaluating the similarity of the two partitions of $P^{(S)}$ and $P^{(F)}$, we can also obtain that $a = 0, b = 0, c = \frac{N(N-1)}{2} = \lambda, d = 0$. From the definitions of the generalized measures in Table 1, it is interesting to find that quite a few generalized measures are not applicable to both of these two cases. Specifically, for the measures Fowlkes-Mallows (FM), Gamma (G), Goodman and Kruskal (GK), Kulczynski (K), McConnaughey (Mc), Pearson (P), Peirce (PE), Sokal and Sneath 1 (SS1), Sokal and Sneath 3 (SS3), the values for both cases turn out to be $\frac{0}{0}$. Also, the two asymmetric measures Wallance 1 (W1) and Wallance 2 (W2) are also not applicable to these two cases. For the other measures, it is interesting to observe that the Adjusted Rand Index (ARI) equals 0, while Baulieu (B) equals 1, which suggests that these two measures do not behave well in this scenario. Among all the generalized measures defined in Table 1, only seven measures: Gower and Legendre (GL), Hamann (H), Jaccard (J), RAND (R), Russel and Rao (RR), Rogers and Tanimoto (RT), Sokal and Sneath 2 (SS2), attain their minimum possible values for this scenario.

3) PROPERTY 4b (P4b): THE MOST UNCERTAIN COMPLEMENTARY PAIR

Besides identifying the most discriminative complement pair, it is also interesting to investigate the performance of these generalized measures when applied to the most uncertain complementary pair. It is straightforward to see that among different cases, $\mathcal{M}_U = 0.5_{N\times N} + 0.5E$ corresponds to the most uncertain one since in this case the probability of each point pair to be in the same cluster is all 0.5. It is also interesting to see that the corresponding value of its complement as defined in Eq. (12) is also 0.5.

We can readily identify these measures with a simple calculation based on Eq. (5) for the case of $\mathcal{M}^{(P)} = \mathcal{M}^{(Q)} = \mathcal{M}_U$, which give $a = b = c = d = \frac{1}{4}\frac{N(N-1)}{2} = \frac{1}{4}\lambda$. Interestingly, we find that two-thirds of the generalized measures attain their mid-point values, which equally divide their ranges between uncertain pairs and certain pairs. This suggests that this property is also considered for measure design. On the other hand, the remaining seven measures, including Gower and Legendre (GL), Jaccard (J), Russel and Rao (RR), Rogers and Tanimoto (RT), Sokal and Sneath 2 (SS2), Sokal and Sneath 3 (SS3), do not have this property. Their similarity values in this scenario and their ranges are listed as follows: GL, $\frac{2}{3}$, [0, 1]; J, $\frac{1}{3}$, [0, 1]; RR, $\frac{1}{4}$, [0, 1]; RT, $\frac{1}{3}$, [0, 1]; SS2, $\frac{1}{5}$, [0, 1]; SS3, $\frac{1}{4}$, [0, 1]. These measures thus do not have this property.

4) PROPERTY 4c (P4c): COMPLEMENTARY PARTITION PAIR

Given a partition P (with co-association matrix $M^{(P)}$) and its complement $E - M^{(P)} + 1$, it is straightforward to observe, based on the binary entries of their respective matrices, that they disagree with each other in terms of the probability of each point pair to be in the same cluster. Although the complement matrix might not necessarily be derived from a valid exclusive hard partition (or a fuzzy ensemble), it is interesting to observe that most of the measures attain their minimum similarity values in this scenario. Thus, the measures that satisfy this criterion should be more well-behaved. For $M^{(P)}$ and $E - M^{(P)} + 1$, we obtain $a = d = 0$ from Eq. (5). On the other hand, we can readily observe three measures: Adjusted Rand Index (ARI), Baulieu (B), and Pearson (P), that do not satisfy this criterion. Specifically, we have

$$\begin{aligned} ARI(M^{(P)}, E - M^{(P)} + 1) &= -\frac{2bc}{b^2 + c^2} \\ B(M^{(P)}, E - M^{(P)} + 1) &= -\frac{(b-c)^2}{(b+c)^2} \\ P(M^{(P)}, E - M^{(P)} + 1) &= -\frac{1}{bc} \end{aligned} \tag{16}$$

However, we find that ARI and B attain their minimum values (-1 and 0, respectively) for the special case $b = c$. In addition, it is notable to observe that all these three

measures attain similarity values below their baseline for complementary partition pairs.

E. PROPERTY 5 (P5): MEASURE OF SELF-SIMILARITY

Given two identical partitions, regardless of their distributions, well-behaved measures should attain their maximum similarity values. On the other hand, for two identical ensembles, the similarity values are not necessarily equal to the maximum values. Viewing this problem from a probabilistic perspective, the self-similarity measure of a partition corresponds to the case where both inputs of the similarity function are the same binary co-association matrix, while that of an ensemble is one where both inputs are the same consensus matrix whose entries signify uncertainty. Let us consider a simple example with one co-association matrix $M = [1\ 1\ 0; 1\ 1\ 0; 0\ 0\ 1]$ and one consensus matrix $\mathcal{M} = [1\ 0.5\ 0.5; 0.5\ 1\ 0.5; 0.5\ 0.5\ 1]$. Intuitively, well-behaved measures should correspond to the case where $sim(M, M)$ is the maximum similarity value. However, for the consensus matrix, $sim(\mathcal{M}, \mathcal{M})$ is not necessarily the maximum value. In fact, we can observe that none of the measures attain the maximum value for the case of the most uncertain complementary pair $sim(\mathcal{M}_U, \mathcal{M}_U)$, where $\mathcal{M}_U = 0.5_{N\times N} + 0.5E$. A simple study of this problem for generalized Adjusted Rand Index is also performed in our recent work [21]. We shall study this self-similarity issue for the fully-connected partition $P^{(F)}$ and the singleton partition $P^{(S)}$ below. The performance of these measures on ensembles will be studied in the experiment section.

1) PROPERTY 5a (P5a): SELF-SIMILARITY FOR THE FULLY-CONNECTED PARTITION $sim(P^{(F)}, P^{(F)})$

Based on Eq. (5), we obtain $a = \frac{N(N-1)}{2}, b = c = d = 0$. From the definitions in Table 1, we can readily observe that ARI, GK, GL, P, PE, SS1, and SS3 are not applicable (division by 0), and the other measures all attain their maximum values.

2) PROPERTY 5b (P5b): SELF-SIMILARITY FOR THE SINGLETON PARTITION, $sim(P^{(S)}, P^{(S)})$

Based on Eq. (5), we obtain $a = b = c = 0, d = \frac{N(N-1)}{2}$. From the definitions in Table 1, we can readily observe that only B, GL, H, RAND, and RT attain their maximum values. Notably, RR equals 0 in this case. The other measures are not applicable (division by 0).

F. SUMMARY

A summary of the proposed comparison properties is provided in Table 2, with brief descriptions and desirable outputs. Also, a summary of the different behaviors of the generalized measures is provided in Table 3. For each column corresponding to a desirable property, 'Y'/'N' indicates whether a particular similarity measure possess this property or not, 'Min'/'Max' represents the maximum/minimum range value, and 'Baseline Value' is the baseline value of each generalized measure. Note that for properties P4a1, P4a2 and P4b, the measure values are also shown in brackets following Y/N to provide more details. Also, 'NA' indicates that a measure is not applicable for a specific property due to division by zero. The fraction of measures with the desirable property is also included in the last row.

From Table 3 we can observe that each property is possessed by at least a few generalized measures. However, it is important to observe that there is not a single measure which possesses all the desirable properties. It is also interesting to note that some measures have similar performance among these properties in spite of their different formulations. Specifically, two groups of measures, G/GK and FM/K, attain the same performance.

These properties can serve as important criteria for the design and selection of evaluation measures. Specifically,

- Property P1 (Symmetry) is important in evaluating pairwise similarity between ensembles. It will be natural to assign the same similarity value between two clusterings or ensembles irrespective of their orders;

TABLE 2. Summary of different properties proposed in this paper.

Property	Brief Descriptions	Desirable Output
P1	Symmetry	$sim(\mathcal{M}^{(P)}, \mathcal{M}^{(Q)}) = sim(\mathcal{M}^{(Q)}, \mathcal{M}^{(P)})$
P2	Possibility of distributed computation	Included
P3	Detection of Uncorrelatedness (DoU)	Included
P3a	DoU between two random partitions	Baseline values
P3b	DoU between two random ensembles	Baseline values
P4	Measure of complementarity	Included
P4a1	$sim(1_{N\times N}, E)$	Minimum values
P4a2	$sim(E, 1_{N\times N})$	Minimum values
P4b	$sim(\mathcal{M}_U, \mathcal{M}_U)$	Baseline values
P4c	$sim(M^{(P)}, E - M^{(P)} + 1)$	Minimum values
P5	Measure of self-similarity	Included
P5a	$sim(1_{N\times N}, E)$	Maximum values
P5b	$sim(E, E)$	Maximum values

TABLE 3. Comparison of generalized measures based on different properties.

Measures	Range	Base	P1	P2	P3a	P3b	P4a1	P4a2	P4b	P4c	P5a	P5b
Desirable			Y	Y	Y	Y	Min	Min	Base	Min	Max	Max
ARI	[-1, 1]	0	Y	Y	Y	Y	N (0)	N (0)	Y (0)	N	NA	NA
B	[0, 1]	0.5	Y	Y	N	N	N (1)	N (1)	Y (0.5)	N	Y	Y
CZ	[0, 1]	0.5	Y	Y	N	N	Y (0)	Y (0)	Y (0.5)	Y	Y	NA
FM	[0, 1]	0.5	Y	Y	N	N	NA	NA	Y (0.5)	Y	Y	NA
G	[-1, 1]	0	Y	Y	Y	Y	NA	NA	Y (0)	Y	NA	NA
GK	[-1, 1]	0	Y	Y	Y	Y	NA	NA	Y (0)	Y	NA	NA
GL	[0, 1]	0.5	Y	Y	N	N	Y (0)	Y (0)	N (0.67)	Y	Y	Y
H	[-1, 1]	0	Y	Y	N	N	Y (-1)	Y (-1)	Y (0)	Y	Y	Y
J	[0, 1]	0.5	Y	Y	N	N	Y (0)	Y (0)	N (0.33)	Y	Y	NA
K	[0, 1]	0.5	Y	Y	N	N	NA	NA	Y (0.5)	Y	Y	NA
MC	[-1, 1]	0	Y	Y	N	N	NA	NA	Y (0)	Y	Y	NA
P	[-1, 1]	0	Y	Y	Y	Y	NA	NA	Y (0)	N	NA	NA
PE	[-1, 1]	0	N	Y	Y	Y	NA	NA	Y (0)	Y	NA	NA
RAND	[0, 1]	0.5	Y	Y	N	N	Y (0)	Y (0)	Y (0.5)	Y	Y	Y
RR	[0, 1]	0.5	Y	Y	N	N	Y (0)	Y (0)	N (0.25)	Y	Y	N(0)
RT	[0, 1]	0.5	Y	Y	N	N	Y (0)	Y (0)	N (0.33)	Y	Y	Y
SS1	[0, 1]	0.5	Y	Y	Y	Y	NA	NA	Y (0.5)	Y	NA	NA
SS2	[0, 1]	0.5	Y	Y	N	N	Y (0)	Y (0)	N (0.2)	Y	Y	NA
SS3	[0, 1]	0.5	Y	Y	N	N	NA	NA	N (0.25)	Y	NA	NA
W1	[0, 1]	0.5	N	Y	N	N	Y (0)	NA	Y (0.5)	Y	Y	NA
W2	[0, 1]	0.5	N	Y	N	N	NA	Y (0)	Y (0.5)	Y	Y	NA
Proportion			$\frac{18}{21}$	$\frac{21}{21}$	$\frac{6}{21}$	$\frac{6}{21}$	$\frac{9}{21}$	$\frac{9}{21}$	$\frac{15}{21}$	$\frac{18}{21}$	$\frac{14}{21}$	$\frac{5}{21}$

- Property P2 (Possibility of distributed computation) will facilitate the distributed computation of the measure values;
- Property P3 (Detection of Uncorrelatedness (DoU)) is well recognized in previous works in distinguishing meaningful partitions/ensembles from random partitions, while in this paper we extend this notion to the scenario of ensembles;
- Property P4 (Measure of complementarity) is important in identifying the most discriminative/uncertain complementary pairs;
- Property P5 (Measure of self-similarity) is important in measuring the degree of self-similarity of partitions/ensembles with different levels of uncertainty.

These observations might also provide an explanation for a number of problems discussed in previous works but not yet solved, which include: (1) Why Fowlkes-Mallows (FM) tends to vary within $[0.6, 1]$ and Rand (R) tends to vary within $[0.5, 0.95]$ for partitions with unbalanced data point distributions [24]; (2) Under what scenarios will negative values be observed for these generalized measures [8], [24]; (3) How these measures perform when applied under different conditions, e.g., between two random partitions (or two random ensembles)(Property 4), or between pairs of different consensus matrices (Property 5)?

V. EXPERIMENTS

We have conducted a number of experiments to investigate the properties of the generalized measures, as well as some of their applications. More specifically, we will mainly investigate the property, Detection of Uncorrelatedness (DoU), for different generalized measures. The dependence of DoU on different variants is consequently presented. We then conduct further experiments to compare these generalized measures based on a number of public data sets. Application of generalized measures to characterize the diversity of cluster ensembles is also discussed.

A. EXPERIMENTS: DETECTION OF UNCORRELATEDNESS (DoU)

The experiments in this subsection are conducted for the following purposes: (i) to verify the detection of uncorrelatedness property for different generalized measures; (ii) to investigate the effect of different factors, such as the number of clustering solutions L, the number of points N and the number of clusters K. In previous sections, we use vague descriptions such as "if N is sufficiently large". These experiments can shed some light on issues such as "What value of N is large enough?".

We first compare two groups of measures on the detection of uncorrelatedness property. Specifically, Group 1 contains six measures, which include 'ARI', 'G', 'GK', 'P', 'PE' and 'SS1'. Group 2 contains the other measures. We generate two random ensembles according to the following specification: the number of data points $N = 100$, the number of partitions for each ensemble $L = 10$, the maximum possible number of clusters $K = 5$ and the number of repeated trials $T = 20$. Figure 1 shows the absolute residual error values of different similarity values after their baseline values are subtracted. From this figure, we can observe that the mean values of all six measures in Group 1 are close to zero, while those in Group 2 are quite different from zero. The experiment results agree well with our analysis in the previous sections.

We further investigate which parameters affect the results corresponding to this property of DoU: the number of data points N, the number of partitions in ensembles L, or the maximum number of clusters K. We only study the performance of the measures in Group 1 for different parameter

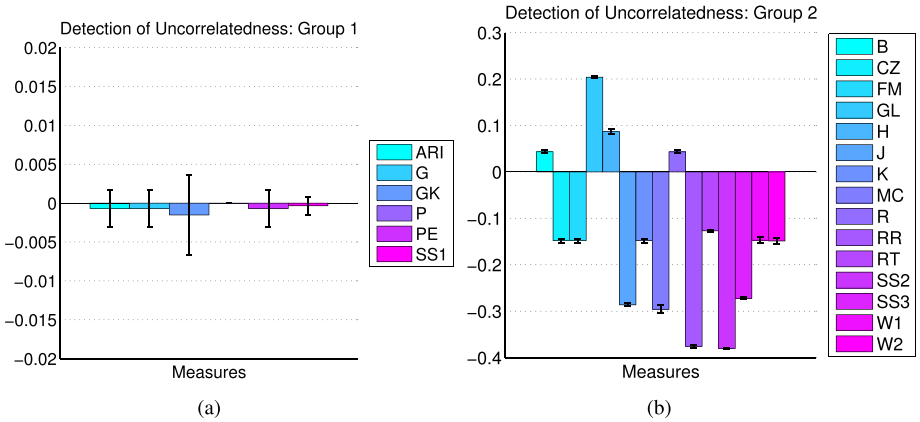


FIGURE 1. Detection of uncorrelatedness for two random ensembles. Group 1 includes six measures which have the desirable property(close to zero), while measures in Group 2 do not have the property (different from zero). (a) Group 1. (b) Group 2.

settings, ranging from $N_{min} = 100, L_{min} = 10, K_{min} = 5$ to $N_{max} = 1000, L_{max} = 100, K_{max} = 20$. To study the effect of a single parameter, we fix the other two parameters. The number of repeated trials is set to $T = 20$. Results of six different scenarios are shown in Figure 2. From this figure, we can observe that (i) the values of all six measures decrease when N increases (Figure 2(a) and Figure 2(b)) and when L increases (Figure 2(e) and Figure 2(f)), but are insensitive to K (Figure 2(c) and Figure 2(d)); (ii) the GK curves are higher than others; (iii) the ARI, G, and PE curves are very close to each other, and (iv) the P curve almost overlaps with the horizontal axis. The first observation indicates that the detection of uncorrelatedness property for well-behaved measures become more obvious with a larger number of data points and/or with a larger number of partitions. The last three observations are in good agreement with the definitions of the measures. These differences come from the fact that their numerators converge to zero while they have different denominators. Specifically, the denominators of ARI, G, GK, and PE are second-order factors of a, b, c, d, while that of P is fourth-order. Moreover, GK has the smallest denominator.

B. FURTHER EXPERIMENTS

We conduct further experiments using nine well-known public data sets from the UCI machine learning repository,[1] including UCI-Breast-Cancer-Wisconsin, UCI-BCW, UCI-Chart, UCI-Glass, UCI-Iris, UCI-Image-Segmentation, UCI-Pima, UCI-Vehicle, and UCI-Wine. These have been used to evaluate the performance of different previous clustering and cluster ensemble techniques.

1) COMPARISON BETWEEN A PARTITION AND A SIMILARITY MATRIX

An intuitive application of the generalized measures is to evaluate similarity between a partition P and an ensemble $\mathcal{Q}$ (with consensus matrix $\mathcal{M}^{(\mathcal{Q})}$). Application of two generalized Adjusted Rand indices (ARImp and ARImm) under this scenario were explored in our recent works [20], [21]. Note that for other measures but not including the pair-counting similarity measures, researchers tend to use the mean value of these measures between the partition P and each partition $Q^{(l)}$, i.e., $sim(P, \mathcal{Q}) = \frac{1}{L}\sum sim(P, Q^{(l)})$ for the same purpose [5], [10]. Thus, a study of the relationship between the results based on $sim(P, \mathcal{M}^{(\mathcal{Q})}))$ and the traditional method (i.e., $sim(P, \mathcal{M}^{(\mathcal{Q})})) = \frac{1}{L}\sum sim(P, Q^{(l)}))$ is of great interest. For each UCI data set, we run Kmeans with the true number of clusters to obtain an initial clustering solution in each trial. Next, we generate a random value $L \in [25, 250]$, and run Kmeans L times to obtain a corresponding number of partitions. We apply each measure to these partitions using our method and the traditional one respectively, and the mean Pearson correlation coefficient between the two different sets of results averaged across 20 trials are reported in Figure 3. Despite the different characteristics of the data sets, experimental results of most of the measures are highly correlated to those of the traditional method, except for B, GK, W1 and W2. These results suggest that most of our generalized measures can achieve results similar to those of the traditional method, while our approach only requires access to the consensus matrix of the ensemble, without the need to observe each individual partition. This advantage becomes more important when the similarity matrix is not constructed from a partition but directly specified. It is also interesting to note that the values for W1 and W2 are uniformly low across all the datasets. This observation might be due to how these two measures are formulated as follows: (i) their definitions do not include the factor d, which is usually larger than the other three factors; and (ii) their numerators include either b or c only ($a + b$ for W1 and $a + c$ for W2), while most of the other measures have both the b and c factors present in an interchangeable way.

[1] http://archive.ics.uci.edu/ml/datasets.html

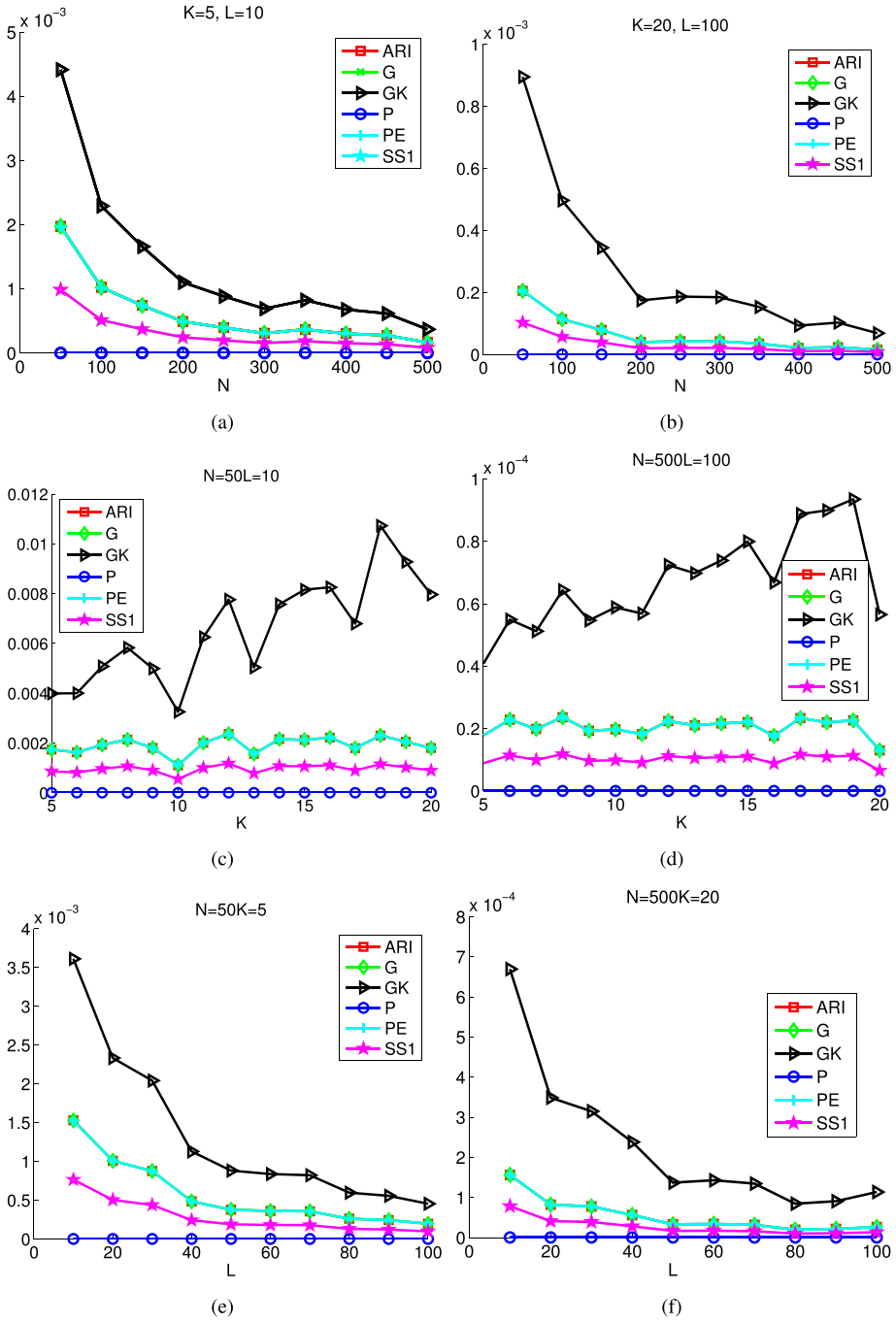


FIGURE 2. Investigation of the effect of different parameters on the detection of uncorrelatedness property: the number of data points N, the number of partitions in ensembles L, and the maximum number of clusters K.

2) OVERALL CORRELATION OF DIFFERENT MEASURE PAIRS IN THE COMPARISON BETWEEN A PARTITION AND A SIMILARITY MATRIX

We have also obtained the pairwise correlation of the 21 generalized measures based on the comparison between a partition and a similarity matrix in the last subsection, which we visualize in Figure 4. Although we only compare these measures for the nine UCI data sets, we can already observe their diverse behavior. It is interesting to note that some measure pairs have quite large correlation, e.g., ARI and G, CZ and J, CZ and FM, SS1 and SS3. On the other hand, H, W1, and W2 appear to be the most different measures compared to the others. We hope that this comparison might provide some help when practitioners choose the desirable subset of measures for their own task. These observations might also be useful when we need to select multiple

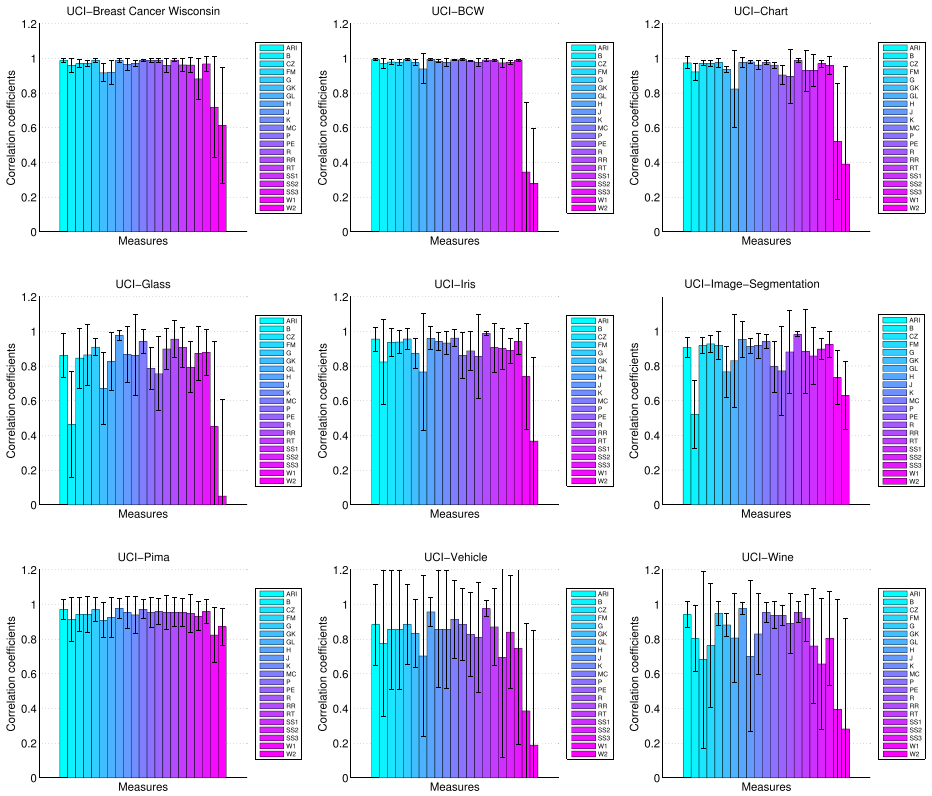


FIGURE 3. Application of generalized measures to characterize the similarity between a partition and an ensemble. Results show correlation coefficient values between our methods and traditional methods.

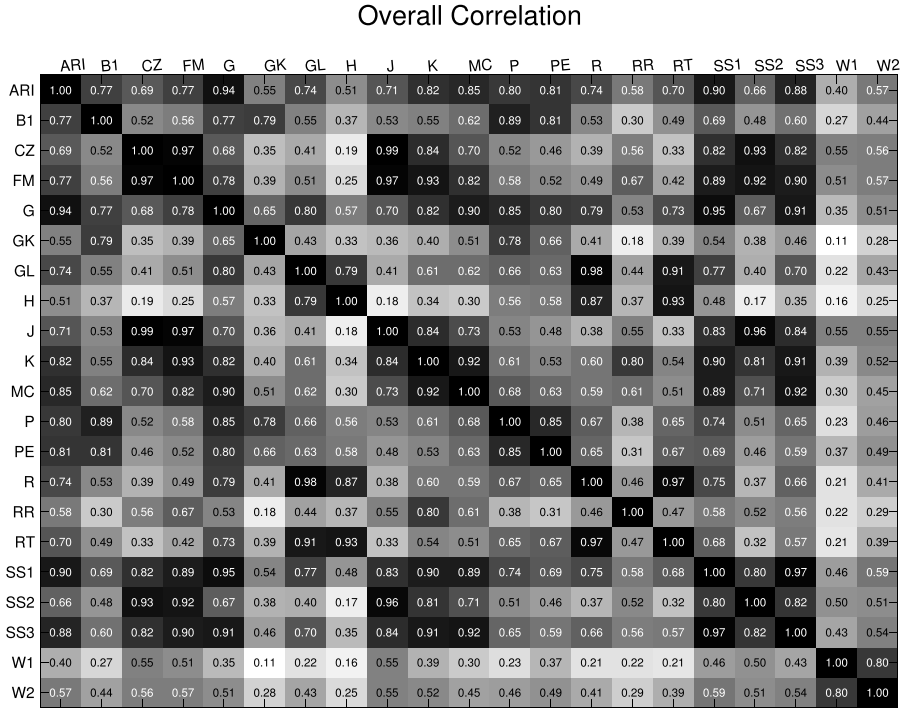


FIGURE 4. Overall pairwise correlation of the 21 generalized measures on nine datasets.

generalized measures to perform a more objective comparison between different clustering (or cluster ensemble) algorithms.

3) MEASURING THE DIVERSITY OF CLUSTER ENSEMBLES

Previous works suggest that the quality of a cluster ensemble is related to its diversity [5], [37], [53]. In general,

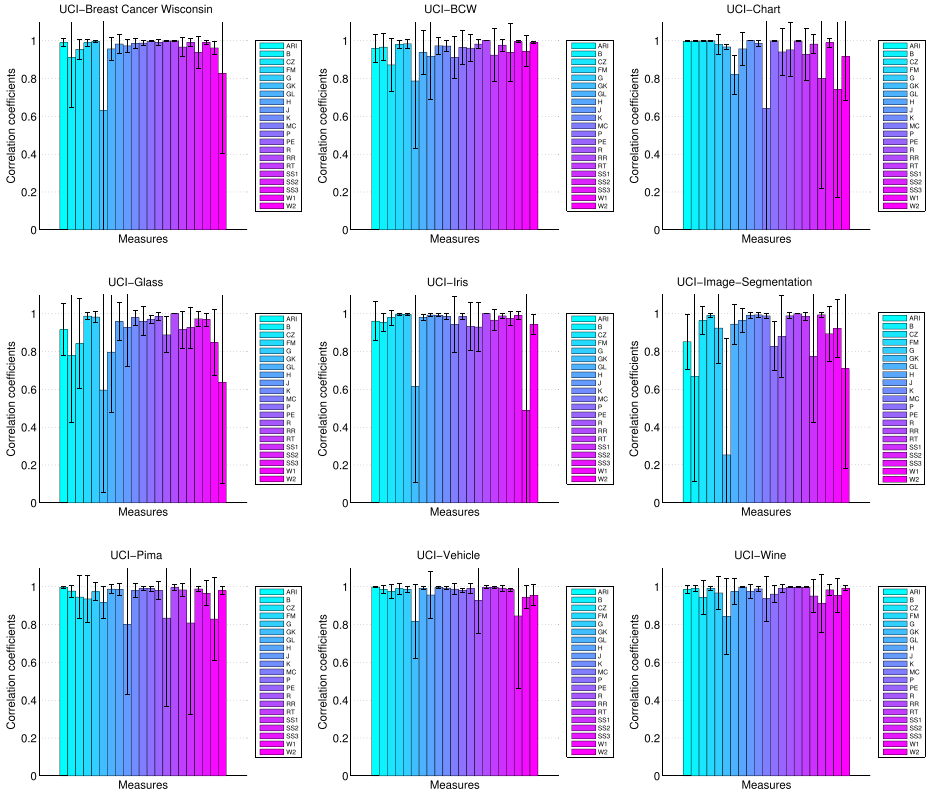


FIGURE 5. Application of generalized measures to characterize the diversity of cluster ensembles. Results show correlation coefficient values between our methods and traditional methods.

traditional methods use pairwise similarity between clustering solutions to measure the diversity of a cluster ensemble (i.e., $\frac{1}{\lambda}\sum_{i=1}^{N}\sum_{j=i+1}^{N} sim(P^{(i)}, P^{(j)})$) such as Pairwise Normalized Mutual Information (PNMI) [5], [37]. In this application, motivated by the traditional method, we use $sim(\mathcal{M}^{(P)}, \mathcal{M}^{(P)})$ between ensembles, rather than partition pairs, to approximate the traditional method. The relation between these two methods is then discussed.

Specifically, we generate 600 clustering solutions for different UCI data sets with the Kmeans algorithm. We use different cluster numbers sampled at random, and group these solutions into three classes using the spectral clustering algorithm as performed in [5]. We refer to these three classes as the small cluster class, the medium cluster class, and the large cluster class according to their sizes. Initially, we add the small cluster class into a base group, and compute the similarity between the clustering solutions in the base group using the different generalized measures. Then we divide the medium cluster class into four different groups at random, and add these groups to the base group one by one in ascending order of their sizes. The corresponding similarity at this stage is also computed using the different generalized measures. Finally, the large cluster class is also divided and added to the base group, followed by the computation of their corresponding similarity values. In this way, the diversity of the 600 clustering solutions can be investigated under nine different conditions. For each condition, the similarity values are computed with our method and the traditional method, respectively, and the Pearson correlation coefficient between the two different computation results are reported in Figure 5.

Interestingly, except for some measures such as B, GK, W1 and W2 which behave in an unstable way, most of the measure values such as ARI (the 1st column), FM (the 4th column) and G (the 5th column) are highly correlated to the results of the traditional methods. Note that our approach only needs to access the consensus matrix of the ensemble, while the traditional computational method requires access to each individual partition. Thus our approach is more general, and its unique advantage is especially useful when access to the individual clustering solutions is difficult.

C. FURTHER DISCUSSIONS

We have analyzed our proposed properties for generalized pair-counting similarity measures in the scenarios of partitions and of cluster ensembles from both the perspectives of theoretical analysis and experimental study. As can be seen, each proposed property has its merit, and is possessed by a number of popular measures. Notably, we do not aim at discovering a complete set of properties to evaluate these measures. Instead, we propose a number of important properties to investigate these measures, especially in the scenario of cluster ensembles. These properties can thus serve as important criteria for the design and selection of evaluation

measures for clustering solutions. In general, there is no single measure which possesses all the desirable properties. Thus, we do not incline to recommend any particular measure for evaluation. Instead, using multiple measures for clustering evaluation is more reasonable and less biased. Therefore, it is important to select a number of diverse measures. In addition, our analysis and experimental results discover measures that have a high correlation with each other. For example, we can find measure pairs which are highly correlated from Figure 4, e.g., ARI vs. G (0.94), CZ vs. FM (0.97), CZ vs. J (0.99), CZ vs. SS2 (0.93), FM vs. J (0.97), FM vs. K (0.93), FM vs.SS2 (0.92), FM vs. SS3 (0.90), G vs. MC (0.90), GL vs. R (0.98), GL vs. RT (0.91), H vs. RT (0.93), J vs. SS2 (0.96), K vs. MC (0.92), K vs. SS1 (0.90), K vs. SS3 (0.91), MC vs. SS3 (0.92), R vs. RT (0.97), and SS1 vs. SS3 (0.97).

An interesting extension of generalized similarity measures to be explored is the case of data labels (or similarity matrices) with missing values. This kind of problems can be dealt with using two different methods: 1) prediction-based methods: we can predict the missing values based on other related entries in the matrices. 2) dropout-based methods: we can simply remove the missing entries for both matrices, and revise the normalized factors in related equations, e.g., (4).

VI. CONCLUSION

In this paper, we have compared 21 pair-counting similarity measures in a generalized setting based on both single clustering solutions and cluster ensemble results, and analyzed their desirable properties from both the perspectives of theoretical analysis and experimental study. We identify their different behaviors and their correlations in different scenarios. It is interesting to observe that each property is possessed by at least a few generalized measures. Notably, some measures have similar performance with regard to these properties in spite of their different formulations. These properties can also serve as important criteria for the design and selection of evaluation measures. We have also performed a number of experiments to verify the comparison criteria, and to demonstrate the performance of the different generalized measures in practical applications.

REFERENCES

[1] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Proc. 2nd Eur. Conf. Comput. Learn. Theory*, 1995, pp. 23–37.

[2] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[3] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.

[4] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004. p. 36.

[5] X. Z. Fern and W. Lin, "Cluster ensemble selection," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2008, pp. 787–797.

[6] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Mach. Learn.*, vol. 52, nos. 1–2, pp. 91–118, Jul. 2003.

[7] A. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835–850, Jun. 2005.

[8] L. I. Kuncheva and D. P. Vetrov, "Evaluation of stability of K-means cluster ensembles with respect to random initialization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1798–1808, Nov. 2006.

[9] H. G. Ayad and M. S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 160–173, Jan. 2008.

[10] J. Azimi and X. Fern, "Adaptive cluster ensemble selection," in *Proc. 21st Int. Joint Conf. Artif. Intell.*, 2009, pp. 992–997.

[11] N. Iam-On, T. Boongoen, and S. Garrett, "LCE: A link-based cluster ensemble method for improved gene expression data analysis," *Bioinformatics*, vol. 26, no. 12, pp. 1513–1519, 2010.

[12] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2396–2409, Dec. 2011.

[13] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Trans. Knowl. Discovery From Data*, vol. 1, no. 1, p. 4, 2007.

[14] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[15] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[16] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *J. Amer. Stat. Assoc.*, vol. 78, no. 383, pp. 553–569, 1983.

[17] P. Jaccard, "Etude de la distribution florale dans une portion des Alpes et du Jura," *Bull. Soc. Vaudoise Sci. Naturelles*, vol. 37, no. 142, pp. 547–579, 1901.

[18] A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko, "On similarity indices and correction for chance agreement," *J. Classification*, vol. 23, no. 2, pp. 301–313, 2006.

[19] R. J. G. B. Campello, "A fuzzy extension of the rand index and other related indexes for clustering and classification assessment," *Pattern Recognit. Lett.*, vol. 28, no. 7, pp. 833–841, May 2007.

[20] S. Zhang and H.-S. Wong, "ARImp: A generalized adjusted rand index for cluster ensembles," in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, Istanbul, Turkey, Aug. 2010, pp. 778–781.

[21] S. Zhang, H.-S. Wong, and Y. Shen, "Generalized adjusted rand indices for cluster ensembles," *Pattern Recognit.*, vol. 45, no. 6, pp. 2214–2226, Jun. 2012.

[22] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on Web-page clustering," in *Proc. Workshop Artif. Intell. Web Search (AAAI)*, 2000, pp. 58–64.

[23] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proc. KDD Workshop Text Mining*, vol. 400. Boston, MA, USA, 2000, pp. 525–526.

[24] M. Meilă, "Comparing clusterings—An information based distance," *J. Multivariate Anal.*, vol. 98, no. 5, pp. 873–895, May 2007.

[25] N. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1073–1080.

[26] P. Luo, H. Xiong, G. Zhan, J. Wu, and Z. Shi, "Information-theoretic distance measures for clustering validation: Generalization and normalization," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1249–1262, Sep. 2009.

[27] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 1, pp. 2837–2854, Oct. 2010.

[28] S. Van Dongen, "Performance criteria for graph clustering and markov cluster experiments," Centrum Wiskunde & Inform., Amsterdam, The Netherlands, Tech. Rep. INSR0012, 2000.

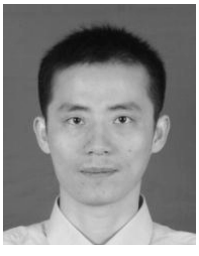
[29] M. Meilă and D. Heckerman, "An experimental comparison of model-based clustering methods," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 9–29, Jan. 2001.

[30] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 1999, pp. 16–22.

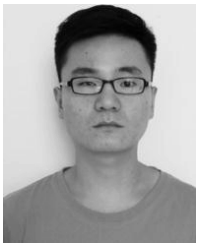
[31] D. T. Anderson, J. C. Bezdek, M. Popescu, and J. M. Keller, "Comparing fuzzy, probabilistic, and possibilistic partitions," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 5, pp. 906–918, Oct. 2010.

[32] X. He, C. H. Q. Ding, H. Zha, and H. D. Simon, "Automatic topic identification using webpage clustering," in *Proc. IEEE Int. Conf. Data Mining*, Nov./Dec. 2001, pp. 195–202.

[33] J. Neville, M. Adler, and D. Jensen, "Clustering relational data using attribute and link information," in *Proc. IJCAI Text Mining Link Anal. Workshop*, 2003, pp. 9–15.
[34] P. Carrington, J. Scott, and S. Wasserman, *Models and Methods in Social Network Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[35] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275–1283, Jul. 2003.
[36] A. Schlicker, F. Domingues, J. Rahnenführer, and T. Lengauer, "A new measure for functional similarity of gene products based on gene ontology," *BMC Bioinformatics*, vol. 7, no. 1, p. 302, 2006.
[37] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 186–193.
[38] J. Wu, S. Zhu, H. Xiong, J. Chen, and J. Zhu, "Adapting the right measures for pattern discovery: A unified view," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 1203–1214, Aug. 2012.
[39] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.
[40] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for K-means clustering," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 877–886.
[41] F. B. Baulieu, "A classification of presence/absence based dissimilarity coefficients," *J. Classification*, vol. 6, no. 1, pp. 233–246, Dec. 1989.
[42] J. Czekanowski, "Coefficient of Racial Likeness und Durchschnittliche Differenz," *Anothropologidcher*, vol. 14, pp. 227–249, Jan. 1932.
[43] L. Goodman and W. H. Kruskal, "Measures of association for cross classifications," *J. Amer. Stat. Assoc.*, vol. 49, no. 268, pp. 732–764, 1954.
[44] J. C. Gower and P. Legendre, "Metric and Euclidean properties of dissimilarity coefficients," *J. Classification*, vol. 3, no. 1, pp. 5–48, 1986.
[45] U. Hamann, "Weiteres über merkmalsbestand und verwandtschaftsbeziehungen der 'Farinosae,'" *Willdenowia*, vol. 3, no. 1, pp. 169–207, 1962.
[46] S. Kulczynski, "Die pflanzenassociationen der pienenen," *Bull. Int. Acad. Polonaise Sci. Lett., Classe Sci. Math. Naturelles B*, vol. 2, pp. 57–203, 1927.
[47] B. H. McConnaughey and L. P. Laut, *The Determination and Analysis of Plankton Communities*. Indonesia: Lembaga Penelitian Laut, 1964.
[48] C. S. Peirce, "The numerical measure of the success of predictions," *Science*, vol. 4, no. 93, pp. 453–454, Nov. 1884.
[49] P. Russel and T. Rao, "On habitat and association of species of anopheline larvae in south-eastern madras," *J. Malaria Inst. India*, vol. 3, no. 1, pp. 153–178, 1940.
[50] D. J. Rogers and T. Tanimoto, "A computer program for classifying plants," *Science*, vol. 132, no. 3434, pp. 1115–1118, 1960.
[51] R. Sokal and P. Sneath, *Principles of Numerical Taxonomy*. San Francisco, CA, USA: Freeman, 1963.
[52] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, 2007, pp. 410–420.
[53] L. I. Kuncheva, S. T. Hadjitodorov, and L. P. Todorova, "Experimental comparison of cluster ensemble methods," in *Proc. 9th Int. Conf. Inf. Fusion*, Jul. 2006, pp. 1–7.



SHAOHONG ZHANG received the Ph.D. degree from the Department of Computer Science, City University of Hong Kong. He was a Post-Doctoral Fellow with the Department of Computer Science, City University of Hong Kong. He is currently an Associate Professor with the Department of Computer Science, Guangzhou University. His research interests include pattern recognition, data mining, and bioinformatics.

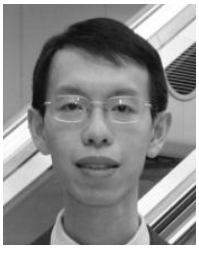


ZONGBAO YANG received the B.S. degree from Guangdong Medical University, Dongguan, China, in 2016. He is currently pursuing the M.S. degree with Guangzhou University, Guangzhou, China. His research interests include data mining and citation network.

XIAOFEI XING (M'17) received the Ph.D. degree in computer science from Central South University, China, in 2012. He has been a Visiting Research Fellow with the University of Tsukuba, Japan. He is an Assistant Professor with the Department of Computer Science, Guangzhou University. His research interests include modeling and performance evaluation in wireless sensor networks and mobile computing. He is a member of the China Computer Federation.

YING GAO received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2002. He is currently a Professor with the School of Computer Science and Educational Software, Guangzhou University, Guangzhou. His main research interests include intelligent optimization algorithms, pattern recognition, and signal processing.

DONGQING XIE was born in Hunan, China, in 1965. He received the M.S. degree from Xidian University in 1988 and Ph.D. degree from Hunan University in 1999. He has been a Professor with Hunan University since 2001. He is currently a Professor with the Department of Computer Science, Guangzhou University. His research interest includes pattern recognition, information security, algorithm analysis and design.



HAU-SAN WONG was a Research Associate with the School of Electrical and Information Engineering, The University of Sydney, and a Post-Doctoral Teaching Fellow with the Department of Computer Science, Hong Kong Baptist University. He is currently an Associate Professor with the Department of Computer Science, City University of Hong Kong. His research interests include bioinformatics and machine learning.