Pattern Recognition 45 (2012) 2214–2226

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Generalized Adjusted Rand Indices for cluster ensembles

Shaohong Zhang, Hau-San Wong*, Ying Shen

Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Hong Kong, PR China

ARTICLE INFO

Article history:
Received 17 June 2011
Received in revised form
1 September 2011
Accepted 24 November 2011
Available online 3 December 2011

Keywords:
Cluster ensembles
Clustering evaluation
Adjusted Rand Index

ABSTRACT

In this paper, Adjusted Rand Index (ARI) is generalized to two new measures based on matrix comparison: (i) Adjusted Rand Index between a similarity matrix and a cluster partition (ARImp), to evaluate the consistency of a set of clustering solutions with their corresponding consensus matrix in a cluster ensemble, and (ii) Adjusted Rand Index between similarity matrices (ARImm), to evaluate the consistency between two similarity matrices. Desirable properties of ARI are preserved in the two new measures, and new properties are discussed. These properties include: (i) detection of uncorrelatedness; (ii) computation of ARImp/ARImm in a distributed environment; and (iii) characterization of the degree of uncertainty of a consensus matrix. All of these properties are investigated from both the perspectives of theoretical analysis and experimental validation. We have also performed a number of experiments to show the usefulness and effectiveness of the two proposed measures in practical applications.

1. Introduction

Clustering is an important approach for organizing vast sets of data in the real world. However, given that a lot of clustering algorithms have been proposed, it is still difficult to select a single one which works well for different data sets. An important technique known as cluster ensemble [1] provides an alternative framework for combining multiple clustering solutions into a more accurate solution. Compared to the individual solutions, cluster ensembles usually provide improved and more stable results.

A cluster ensemble technique can be characterized by two main phases: (i) to generate a number of individual clustering solutions, and (ii) to find a final consensus solution with these clustering solutions. In the first phase, representative methods to generate individual clusterings include (i) clustering with different point subsets of the original data based on different sampling techniques [2,3], (ii) clustering with different feature subsets or feature projection techniques [3–5], (iii) clustering with different initialization conditions (such as different cluster numbers or different initialization seeds) [3,6–10], and (iv) clustering with different algorithms [1,11]. In the second phase, a consensus clustering solution is derived from the individual clusterings based on different methods. Representative methods can be roughly categorized into two types: (i) co-association based methods and (ii) graph mapping based methods. In the first type, a co-association matrix is first derived from each clustering solution, whose entries specify whether a pair of data points belong to the same cluster according to the current clustering solution. Different consensus functions have been used for combining the co-association matrices, in which the consensus matrix is one of the most popular formulations. In particular, each entry of the consensus matrix is the mean of the corresponding entries from all the co-association matrices. Such a consensus matrix can be viewed as a similarity matrix, and thus a final clustering solution can be obtained with any clustering algorithm that work directly on distance/similarity matrices. Cluster ensemble methods based on hierarchical agglomerative clustering algorithms with Single Link (SL), Average Link (AL) or Complete Link (CL) are popular representatives of this type of methods [7]. On the other hand, the consensus matrix can also be viewed as a new data matrix, and the final clustering solution can be obtained with a conventional clustering algorithm [12], such as the well-known k-means algorithm [13]. Another popular type of cluster ensemble methods is to map the individual clusterings to different graphs, and thus the problem can be solved with many different popular graph cut algorithms. Among this category, typical solutions include cluster-based similarity partitioning algorithm (CSPA) [1], Meta-CLustering Algorithm (MCLA) [1], and the Hybrid Bipartite Graph Formulation algorithm (HBGF) [4].

Among the cluster ensemble techniques, one of the most popular approaches to combine multiple clustering solutions is to construct a consensus matrix. The advantages of this approach include: (1) the consensus matrix corresponds to a more stable representation of a partition than an individual clustering solution. (2) We can assign different importance weightings to the different clustering solutions based on prior knowledge.

* Corresponding author. Tel.: +852 34428624.
E-mail addresses: zimzsh@gmail.com (S. Zhang), cshswong@cityu.edu.hk (H.-S. Wong), yingshen3@student.cityu.edu.hk (Y. Shen).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.11.017

(3) Different kinds of clustering techniques can be used to create the ensemble. However, while different measures have been proposed in previous works to characterize the consistency between different clustering solutions in a cluster ensemble, for example Adjusted Rand Index (ARI) [14] and Normalized Mutual Information (NMI) [1], there does not exist a corresponding measure to characterize the consistency between a specific clustering solution and a consensus matrix. In addition, there is a lack of meaningful measures to characterize the consistency between pairwise similarity matrices such as co-association matrices or consensus matrices in cluster ensembles.

It is notable that the requirement to evaluate the consistency between a clustering solution and a consensus matrix, or between two similar matrices, is not limited to the case of cluster ensembles. In real applications, many data sets might be extracted from different perspectives. In general, a lot of data sets include not only vectors of attribute values (referred to as features in conventional clustering approaches) but also pairwise relation information (which is sometimes called similarity). Representative examples could be easily found, such as pairwise linkage analysis in web data mining [15,16], structural relation information in social network analysis [17], and semantic similarity computation in Gene Ontology [18,19]. Recently, clustering based on both the attribute and relation information becomes an important topic, and many novel clustering algorithms have been developed [20,21]. However, to our best knowledge, the issues of how to effectively evaluate the consistency between a clustering solution and a relationship graph, and the similarity between two relationship graphs are still open problems.

In view of the discussion above, we investigate these problems by proposing new measures along the line of ARI and its fuzzy extension [22]. We choose ARI because it can be readily computed based on the pairwise relation matrices, while it is in general difficult to compute NMI in a similar way. Specifically, we propose two new measures: (i) Adjusted Rand Index between a similarity matrix and a cluster partition (ARImp) and (ii) Adjusted Rand Index between two relation matrices (ARImm) for comparing these two matrices. We show that ARImp and ARImm are fuzzy generalization of ARI, and the equivalence between ARI, ARImp and ARImm are proved from the viewpoint of matrix comparison. Desirable properties of ARI are preserved in ARImp and ARImm, and new attractive properties of ARImp and ARImm are discussed and proved. Note that another advantage of cluster ensemble is to facilitate the implementation of clustering in a distributed environment where the raw data cannot be shared among users due to different restrictions on storage, privacy and ownership [1]. Compared to other measures, another significant attractive property of ARImp/ARImm is their ability to enhance the cluster analysis process in the above context. Specifically, ARImp/ARImm compare ensembles based on consensus matrices instead of the individual clustering solutions. We can illustrate the benefit of this property by the following example: given the availability of a number of clusterings from different companies based on the same set of data, a third-party organization (such as a government department) might need to collect these data to perform statistical analysis. However, the organization cannot release the individual clusterings to the companies, since these individual solutions might contain trade secrets. On the other hand, a consensus ensemble might be released. In this case, with the availability of only the ensemble result, each company cannot evaluate its own clustering based on traditional measures. However, ARImp can effectively deal with this problem. Similar examples can be found for ARImm. For example, assume that there are two government departments collecting different individual clusterings for the same set of data, and only two consensus results are released. In this case ARImm can be used to evaluate the similarity based on these two consensus ensembles alone, rather than relying on the different individual clusterings which cannot be shared with each other. More application examples of ARImp and ARImm are also presented in the experiment section.

1.1. Contributions of this paper

The main contribution of this paper is the formulation of two new measures, ARImp and ARImm, which allow the effective comparison between clustering solutions and consensus matrices, and that between consensus matrices respectively. We investigate these two new measures from both the perspectives of theoretical analysis and experimental validation. In addition, we provide a number of application examples for ARImp and ARImm, which show the effectiveness of these two new proposed measures.

1.2. Organization of this paper

The rest of the paper is organized as follows. Section 2 introduces previous works on cluster ensembles. We also describe two popular measures for clustering solution evaluation (ARI and NMI), and some previous measures for the comparison between clustering solutions and consensus matrices, and those between consensus matrices. Section 3 describes the two proposed measures ARImp and ARImm. We show that ARImp and ARImm are fuzzy generalization of ARI, and the equivalence between ARI, ARImp and ARImm are proved from the viewpoint of matrix comparison. Desirable properties of ARImp and ARImm are proposed in Section 4. Experimental results are discussed in Section 5. Conclusions are presented in Section 6.

2. Cluster ensembles and related clustering measures

In this section, we provide background information about cluster ensembles, and describe two popular measures, ARI and NMI, for clustering solution evaluation. We also describe some previous measures for the comparison between clustering solutions and consensus matrices, and those between consensus matrices.

2.1. Cluster ensembles and consensus matrix

Specifically, given a data set $X=\{x_i\}_{i=1}^{N}$ with N points, a clustering (or partition) P partitions X into a number of mutually disjoint subsets $\{P_k\}_{k=1}^{K}$ called clusters. The $N\times N$ co-association matrix for the partition P is defined as follows:

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

Given a number of clustering solutions $\{P^{(l)}\}_{l=1}^{L}$, the $N\times N$ consensus matrix for these clustering partitions can be constructed as the average of the individual co-association matrices as follows:

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

It is interesting to note that a consensus matrix can be viewed as a fuzzy generalization of a co-association matrix. An illustration of the consensus ensemble formation process is shown in Fig. 1.

2.2. Clustering measures: Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI)

In this subsection, we introduce two popular measures for comparing different clusterings in the cluster ensemble literature: Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

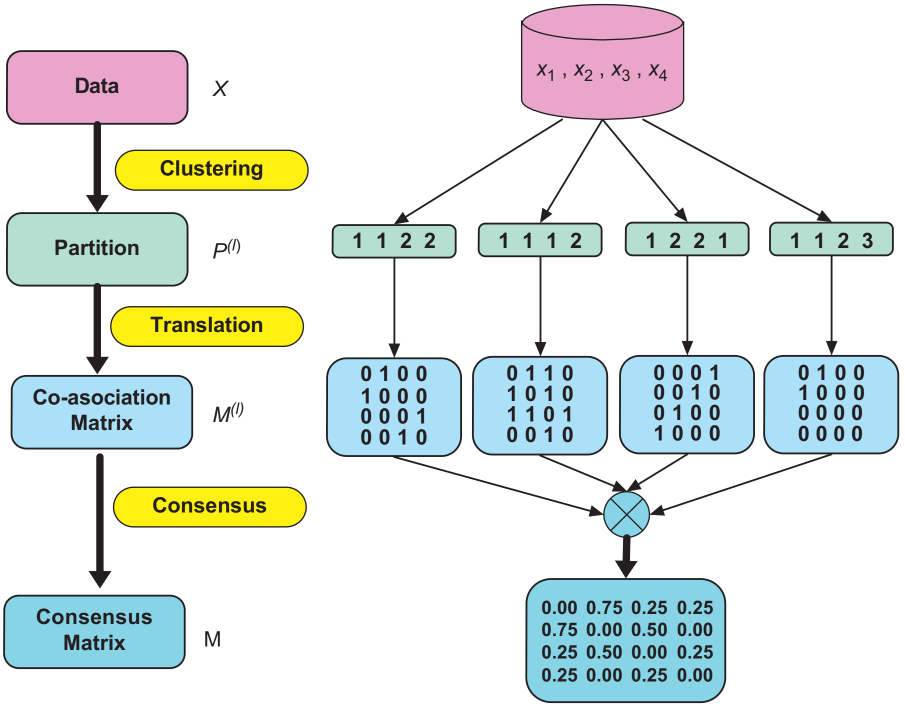

Fig. 1. Illustration of the consensus ensemble formation process. A simple example is shown on the right.

Table 1
The contingency table.

Cluster	Q_1	Q_2	$\cdots$	$Q_{K^{(Q)}}$	$\sum$
P_1	N_{11}	N_{12}	$\cdots$	$N_{1K^{(Q)}}$	$N_{1\cdot}$
P_2	N_{21}	N_{22}	$\cdots$	$N^{(Q)}_{2K}$	$N_{2\cdot}$
.	.	.	$\cdots$	.	.
$P_{K^{(P)}}$	$N_{K^{(P)}1}$	$N_{K^{(P)}2}$	$\cdots$	$N_{K^{(P)}K^{(Q)}}$	$N_{K^{(P)}\cdot}$
$\sum$	$N_{\cdot 1}$	$N_{\cdot 2}$	$\cdots$	$N_{\cdot K^{(Q)}}$	N

Let $P=\{P_1,P_2,\ldots,P_{K^{(P)}}\}$ and $Q=\{Q_1,Q_2,\ldots,Q_{K^{(Q)}}\}$ be two partitions on a data set X with N entities, N_{ij} be the number of entities in cluster P_i in partition P and in cluster Q_j in partition Q, $N_{i\cdot}$ be the number of entities in cluster P_i in partition P, $N_{\cdot j}$ be the number of entities in cluster Q_j in partition Q, the degree of similarity between these two partitions can be characterized using a contingency matrix as shown in Table 1. The Adjusted Rand Index (ARI) [14] is defined as follows:

$$r_0=\sum_{i=1}^{K^{(P)}}\sum_{j=1}^{K^{(Q)}}\binom{N_{ij}}{2},\quad r_1=\sum_{i=1}^{K^{(P)}}\binom{N_{i\cdot}}{2}$$

$$r_2=\sum_{j=1}^{K^{(Q)}}\binom{N_{\cdot j}}{2},\quad r_3=\frac{2r_1r_2}{N(N-1)} \tag{3}$$

$$ARI(P,Q)=\frac{r_0-r_3}{0.5(r_1+r_2)-r_3} \tag{4}$$

where $\binom{n}{k}$ is the binomial coefficient.

Normalized Mutual Information (NMI) [1] is another standard measure for clustering, and it is also extensively used in previous works. Specifically, NMI measures how similar two clustering solutions is based on a normalized version of the mutual information measure, and it is defined for two clustering partitions $P=\{P_1,P_2,\ldots,P_{K^{(P)}}\}$ and $Q=\{Q_1,Q_2,\ldots,Q_{K^{(Q)}}\}$ as follows [1]:

$$NMI(P,Q)=\frac{I(P,Q)}{\sqrt{H(P)H(Q)}} \tag{5}$$

where $I(P,Q)$ is the mutual information between P and Q, and $H(P)$ is the entropy of P.

Average Normalized Mutual Information (ANMI) [1,3] is used to measure the consistency between a set of clustering partitions $\{P^{(l)}\}_{l=1}^{L}$ and a clustering partition Q, which is defined as follows:

$$ANMI(\{P^{(l)}\}_{l=1}^{L},Q)=\frac{1}{L}\sum_{l=1}^{L}NMI(P^{(l)},Q) \tag{6}$$

Diversity of a cluster ensemble is generally agreed to be a contributing factor of a cluster ensemble. In general, the Pairwise Normalized Mutual Information (PNMI) among clustering solutions is used to measure the diversity of a clustering ensemble [23,3]

$$PNMI(\{P^{(l)}\}_{l=1}^{L})=\sum_{i\neq j}NMI(P^{(i)},P^{(j)}) \tag{7}$$

In other words, a lower $PNMI(\{P^{(l)}\}_{l=1}^{L})$ value corresponds to a higher diversity.

2.3. General matrix comparison measures

To our best knowledge, there are no proposed measures to evaluate the similarity between two consensus matrices. As an alternative, since we consider the comparison of two ensembles based on their respective consensus matrices, we use some general matrix comparison measures as references. The most straightforward measure is the Normalized Similarity using the Frobenius norm (NSF) of the difference matrix between two matrices

$$NSF(M^{(P)},M^{(Q)})=1-\frac{1}{N^2}\|M^{(P)}-M^{(Q)}\|_F \tag{8}$$

We choose the kernel alignment similarity measure (KAS) [24] as the second reference, which is widely used to compare the similarity between two kernel matrices [24,25]. For two kernel matrices $M^{(P)}$ and $M^{(Q)}$, the kernel alignment similarity measure (KAS) is defined as follows:

$$KAS(M^{(P)},M^{(Q)}) = \frac{tr(M^{(P)}M^{(Q)^T})}{\sqrt{tr(M^{(P)}M^{(P)^T})tr(M^{(Q)}M^{(Q)^T})}} \quad (9)$$

We choose the scaled Standardized Mantel Statistic (sSMS) as the third reference. The Standardized Mantel Statistic is widely used in the Mantel test [26] to measure the correlation between two matrices [27]. For two square matrices $M^{(P)}$ and $M^{(Q)}$ (the number of both rows and columns equals N), the Standardized Mantel Statistic (SMS) [28] is defined as follows:

$$SMS(M^{(P)},M^{(Q)}) = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\frac{M_{ij}^{(P)}-\overline{M_{ij}^{(P)}}}{s^{(P)}}\right)\left(\frac{M_{ij}^{(Q)}-\overline{M_{ij}^{(Q)}}}{s^{(Q)}}\right)}{d-1} \quad (10)$$

where $\overline{M_{ij}^{(P)}}$ is the mean of $M_{ij}^{(P)}$, $s^{(P)}$ is the standard deviation of $M_{ij}^{(P)}$, and d is the number of entries in the upper triangular portion of either matrix. Since the range of the SMS coefficient is from -1 to $+1$, we use the scaled Standardized Mantel Statistic (sSMS) to transform it to the range [0, 1] as follows:

$$sSMS(M^{(P)},M^{(Q)}) = \frac{1+SMS(M^{(P)},M^{(Q)})}{2} \quad (11)$$

3. Generalized Adjusted Rand Indices for cluster ensembles

In this section, we propose two new generalized measures: (i) Adjusted Rand Index between a similarity matrix and a cluster partition (ARImp), to evaluate the consistency of a set of clustering solutions with their corresponding consensus matrix in a cluster ensemble and (ii) Adjusted Rand Index between two consensus matrix matrices (ARImm), to evaluate the consistency between two similarity matrices.

3.1. Generalized Adjusted Rand Indices for consensus matrices

Based on ARI, we propose the measure ARImp for an $N\times N$ consensus matrix $\mathcal{M}$ and a partition $Q=\{Q_1,Q_2,\ldots,Q_{K^{(Q)}}\}$ (the number of entities in cluster Q_k is denoted by N_k) as follows:

$$s_0 = \sum_{k=1}^{K^{(Q)}}\sum_{i,j\in Q_k,i\neq j}\frac{\mathcal{M}_{ij}}{2}, \quad s_1 = \sum_{i\neq j}\frac{\mathcal{M}_{ij}}{2}$$

$$s_2 = \sum_{k=1}^{K^{(Q)}}\binom{N_k}{2}, \quad s_3 = \frac{2s_1s_2}{N(N-1)} \quad (12)$$

$$ARImp(\mathcal{M},Q) = \frac{s_0-s_3}{0.5(s_1+s_2)-s_3} \quad (13)$$

We also propose the measure ARImm between two similarity matrices $\mathcal{M}^{(P)}$ and $\mathcal{M}^{(Q)}$ as follows:

$$t_0 = \sum_{i\neq j}\frac{\mathcal{M}_{ij}^{(P)}\mathcal{M}_{ij}^{(Q)}}{2}, \quad t_1 = \sum_{i\neq j}\frac{\mathcal{M}_{ij}^{(P)}}{2}$$

$$t_2 = \sum_{i\neq j}\frac{\mathcal{M}_{ij}^{(Q)}}{2}, \quad t_3 = \frac{2t_1t_2}{N(N-1)} \quad (14)$$

$$ARImm(\mathcal{M}^{(P)},\mathcal{M}^{(Q)}) = \frac{t_0-t_3}{0.5(t_1+t_2)-t_3} \quad (15)$$

3.2. Relationship between ARI, ARImp, and ARImm

We now perform comparison between ARI, ARImp, and ARImm. First, we can see that they are all expressed in terms of four factors. To distinguish between the above three measures, we focus on the difference between these factors, or more specifically, on r_0,r_1 and s_0,s_1 between ARI and ARImp, and on s_0,s_1 and t_0,t_1 between ARImp and ARImm. To understand this difference, we first state two lemmas about the factors r_0 and r_1 [14].

Lemma 1. *r_0 in Eq.* (3) *corresponds to the total number of pairs of points which belong to the same cluster in the two partitions.*

Lemma 2. *r_1 in Eq.* (3) *corresponds to the total number of pairs of points which belong to the same cluster in partition P.*

It is interesting to observe that s_0 in Eq. (12) can be regarded as a fuzzy generalization of r_0 in Eq. (3). Note that, for a consensus matrix, an entry, say $\mathcal{M}_{ij}$, represents the probability of point i and point j being in the same cluster. Therefore, s_0 represents the summation of the probabilities of those pairs of points within the same cluster. Similarly, s_1 in Eq. (12) can also be regarded as a fuzzy generalization of r_1 in Eq. (3). In addition, t_0 and t_1 in Eq. (14) can be naturally viewed as a generalization of s_0 and s_1 in Eq. (12). For a certain data set, given two partitions P and Q and their co-association matrices $M^{(P)}$ and $M^{(Q)}$, we would also like to point out the equivalence between these generalized ARI measures and the classic ARI measure.

Proposition 1. *$ARI(P,Q)=ARImp(M^{(P)},Q)=ARImp(P,M^{(Q)})=ARImm(M^{(P)},M^{(Q)})$.*

Proof. Without loss of generality, we only prove (i) $ARI(P,Q)=ARImp(M^{(P)},Q)$ and (ii) $ARImp(P,M^{(Q)})=ARImm(M^{(P)},M^{(Q)})$.

(i) Proof of $ARI(P,Q)=ARImp(M^{(P)},Q)$:

We first focus on a cluster Q_k in the partition Q. The number of pairs of points from Q_k which also belong to the same cluster in partition P is $\sum_{i=1}^{K^{(P)}}\binom{N_{ik}}{2}$. In the case of $ARImp(M^{(P)},Q)$, there are only binary entries in the co-association matrix $M^{(P)}$, where for two points x_i and x_j, the entry $M_{ij}^{(P)}=1$ indicates that they belong to the same cluster in partition P. Thus, $\sum_{i,j\in Q_k,i\neq j}(M_{ij}^{(P)}/2)$ is also equal to the number of pairs of points from Q_k belonging to the same cluster in partition P. Therefore, we obtain $\sum_{i=1}^{K^{(P)}}\binom{N_{ik}}{2}=\sum_{i,j\in Q_k,i\neq j}(M_{ij}^{(P)}/2)$. When summing all these terms across k (from 1 to K_Q), we obtain $\sum_{k=1}^{K^{(Q)}}\sum_{i=1}^{K^{(P)}}\binom{N_{ik}}{2}=\sum_{k=1}^{K^{(Q)}}\sum_{i,j\in Q_k,i\neq j}(M_{ij}^{(P)}/2)$, i.e., $r_0=s_0$. Similarly, we can show that $r_1=s_1$. As a result, $ARI(P,Q)$ is equal to $ARImp(M^{(P)},Q)$.

(ii) Proof of $ARImp(P,M^{(Q)})=ARImm(M^{(P)},M^{(Q)})$:

Considering the computation of ARImp (see Eqs. (12) and (13)) and ARImm (see Eqs. (14) and (15)), we only need to prove the equivalence between s_0 in Eq. (12) and t_0 in Eq. (14). Following Lemma 1, we know that s_0 in Eq. (12) is equal to the total number of pairs of points which belong to the same cluster in the two partitions. Consider a cluster P_k from the partition P, we construct its co-association matrix $M^{(P_k)}$ as follows: $M_{ij}^{(P_k)}=1$ if the two points x_i and x_j belong to P_k and $M_{ij}^{(P_k)}=0$ otherwise. Thus, we can observe that $\sum_{i\neq j}(M_{ij}^{(P_k)}M_{ij}^{(Q)})$ is equal to the number of pairs of points which are in cluster P_k and also belong to the same cluster in partition Q. Since $\sum_{i\neq j}(M_{ij}^{(P)}M_{ij}^{(Q)})=\sum_k\sum_{i\neq j}(M_{ij}^{(P_k)}M_{ij}^{(Q)})$, i.e., $\sum_{i\neq j}(M_{ij}^{(P)}M_{ij}^{(Q)})$ is equal to the number of point pairs that are in the same cluster in both partition P and partition Q, we obtain $t_0=s_0$. In addition, we obtain $s_1=t_1$ and $s_2=t_2$ since they are both equal to the summed probabilities of pairs of points within the same cluster in partition P and partition Q respectively. It then follows that $s_3=t_3$ from Eqs. (12) and (14). Therefore, we can obtain $ARImp(P,M^{(Q)})=ARImm(M^{(P)},M^{(Q)})$. □

4. Properties of generalized Adjusted Rand Indices

In this section, we introduce a number of desirable properties of ARImp and ARImm, including the property "Detection of Uncorrelatedness" inherited from ARI. Specifically, the ARI value between two random partitions is close to zero. We shall prove this property for ARImp and ARImm, i.e., (i) ARImp between a consensus matrix constructed from random partitions and an uncorrelated partition is close to zero and (ii) ARImm between two uncorrelated consensus matrices constructed from two sets of random partitions is close to zero. In this paper, a random partition is obtained when data points are assigned to their clusters at random based on a uniform distribution, i.e., the probability of a point being assigned to one of the clusters is $1/K$, where the number of clusters K is randomly selected from 2 to the maximum number K_{max} with the uniform distribution.

4.1. Detection of uncorrelatedness

Proposition 2. *For a data set with N points, given a consensus matrix $\mathcal{M}$ computed from L random partitions $\{P^{(l)}\}_{l=1}^{L}$, and another partition Q which is uncorrelated to any $P^{(l)}$, we have* $\lim_{N\to\infty}$ *$ARImp(\mathcal{M},Q)=0$.*

Proof. We first focus on the consensus matrix $\mathcal{M}$ which is generated from the average of the individual co-association matrices (Eq. (2)). Given a particular random partition generated with the uniform distribution, say $P^{(l)}$ with $K^{(l)}$ clusters, the probability that the element in its co-association matrix $M^{(l)}$ equals 1 (i.e., for the corresponding pair of points to belong to the same cluster) can be determined as follows:

$$p^{(l)} = p(M_{ij}^{(l)}=1) = \frac{\binom{K^{(l)}}{1}}{K^{(l)}K^{(l)}} = \frac{1}{K^{(l)}}, \quad p(M_{ij}^{(l)}=0) = 1-\frac{1}{K^{(l)}} \tag{16}$$

That is, $p(M_{ij}^{(l)})$ is a Bernoulli distribution.

We then proceed as follows:

$$s_1 = \sum_{i\neq j}\frac{\mathcal{M}_{ij}}{2} = \sum_{i\neq j}\frac{\frac{1}{L}\sum_{l=1}^{L} M_{ij}^{(l)}}{2} = \frac{1}{L}\sum_{l=1}^{L}\sum_{i\neq j}\frac{M_{ij}^{(l)}}{2} = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{N(N-1)}{2}\overline{M_{ij}^{(l)}}\right) \tag{17}$$

where $\overline{M_{ij}^{(l)}}$ is the mean of $M_{ij}^{(l)}$. It is well known that the empirical mean of a set of observed values of a random variable can be approximated using the expectation of the variable, and the expectation of a random variable according to a Bernoulli distribution with probability $p^{(l)}$ is $p^{(l)}$ itself. Thus, $\overline{M_{ij}^{(l)}}$ can be approximated with $p^{(l)}$. Setting $\lambda = N(N-1)/2$, we obtain

$$s_1 = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{N(N-1)}{2}\overline{M_{ij}^{(l)}}\right) \approx \frac{\lambda}{L}\sum_{l=1}^{L} p^{(l)} \tag{18}$$

Similarly, we can obtain

$$s_2 = \sum_{k=1}^{K^{(Q)}}\binom{N_k}{2} = \sum_{i\neq j}\frac{M_{ij}^{(Q)}}{2} = \frac{N(N-1)}{2}\overline{M_{ij}^{(Q)}} \approx \lambda p^{(Q)} \tag{19}$$

In addition,

$$s_3 = \frac{2s_1s_2}{N(N-1)} \approx \frac{\frac{\lambda}{L}(\sum_{l=1}^{L} p^{(l)})\lambda p^{(Q)}}{\lambda} = \frac{\lambda p^{(Q)}}{L}\sum_{l=1}^{L} p^{(l)} \tag{20}$$

On the other hand, we have

$$\begin{aligned} s_0 &= \sum_{k=1}^{K^{(Q)}}\sum_{i,j\in Q_k, i\neq j}\frac{\mathcal{M}_{ij}}{2} = \sum_{i\neq j}\frac{M_{ij}^{(Q)}\mathcal{M}_{ij}}{2} = \sum_{i\neq j}\frac{M_{ij}^{(Q)}\cdot\frac{1}{L}\sum_{l=1}^{L} M_{ij}^{(l)}}{2} \\ &= \frac{1}{2L}\sum_{l=1}^{L}\sum_{i\neq j}[M_{ij}^{(Q)}M_{ij}^{(l)}] = \frac{1}{2L}\sum_{l=1}^{L} N(N-1)\overline{M_{ij}^{(Q)}M_{ij}^{(l)}} \\ &= \frac{N(N-1)}{2L}\sum_{l=1}^{L}\overline{M_{ij}^{(Q)}}\,\overline{M_{ij}^{(l)}} = \frac{\lambda}{L}\overline{M_{ij}^{(Q)}}\sum_{l=1}^{L}\overline{M_{ij}^{(l)}} \approx \frac{\lambda p^{(Q)}}{L}\sum_{l=1}^{L} p^{(l)} \end{aligned} \tag{21}$$

The seventh step uses the fact that partition Q is uncorrelated with the partitions which are used to construct the consensus matrix, and the eighth step follows from the approximation using the expectation when N is large.

Thus, for $ARImp(\mathcal{M},Q)$, from Eq. (13), the numerator is given by

$$s_0 - s_3 \approx \frac{\lambda p^{(Q)}}{L}\sum_{l=1}^{L} p^{(l)} - \frac{\lambda p^{(Q)}}{L}\sum_{l=1}^{L} p^{(l)} = 0 \tag{22}$$

and the denominator is given by

$$\begin{aligned} 0.5(s_1+s_2)-s_3 &\approx 0.5\left(\frac{\lambda}{L}\sum_{l=1}^{L} p^{(l)} + \lambda p^{(Q)}\right) - \frac{\lambda p^{(Q)}}{L}\sum_{l=1}^{L} p^{(l)} \\ &= 0.5\lambda\left(\frac{1}{L}\sum_{l=1}^{L} p^{(l)} + p^{(Q)} - \frac{2p^{(Q)}}{L}\sum_{l=1}^{L} p^{(l)}\right) \end{aligned} \tag{23}$$

In general, the denominator is not equal to zero. Since the numerator is equal to zero, we can conclude that $ARImp(\mathcal{M},Q)=0$. □

Proposition 3. *For a data set with N points, suppose that two consensus matrices $\mathcal{M}^{(P)}$ and $\mathcal{M}^{(Q)}$ are constructed from two uncorrelated random partition sets $\{P^{(l_P)}\}_{l_P=1}^{L_P}$ and $\{Q^{(l_Q)}\}_{l_Q=1}^{L_Q}$ respectively. We have* $\lim_{N\to\infty}$*$ARImm(\mathcal{M}^{(P)},\mathcal{M}^{(Q)})=0$.*

Proof. Using Eq. (21), we have

$$\begin{aligned} t_0 &= \sum_{i\neq j}\frac{\mathcal{M}_{ij}^{(P)}\mathcal{M}_{ij}^{(Q)}}{2} = \sum_{i\neq j}\frac{\frac{1}{L_P}\sum_{l_P=1}^{L_P} M_{ij}^{(l_P)}\cdot\frac{1}{L_Q}\sum_{l_Q=1}^{L_Q} M_{ij}^{(l_Q)}}{2} \\ &= \frac{1}{L_PL_Q}\sum_{l_P=1}^{L_P}\sum_{l_Q=1}^{L_Q}\sum_{i\neq j}\frac{M_{ij}^{(l_P)}M_{ij}^{(l_Q)}}{2} \approx \frac{\lambda}{L_PL_Q}\sum_{l_P=1}^{L_P}\sum_{l_Q=1}^{L_Q}(p^{(l_P)}p^{(l_Q)}) \end{aligned} \tag{24}$$

where the fourth step is obtained in a similar way as the derivation of Eq. (21).

Also, from Eq. (18), we can obtain

$$t_1 \approx \frac{\lambda}{L_P}\sum_{l_P=1}^{L_P} p^{(l_P)}, \quad t_2 \approx \frac{\lambda}{L_Q}\sum_{l_Q=1}^{L_Q} p^{(l_Q)} \tag{25}$$

$$\begin{aligned} t_3 &= \frac{2t_1t_2}{N(N-1)} = \frac{t_1t_2}{\lambda} \approx \frac{\frac{\lambda}{L_P}\sum_{l_P=1}^{L_P} p^{(l_P)}\frac{\lambda}{L_Q}\sum_{l_Q=1}^{L_Q} p^{(l_Q)}}{\lambda} \\ &= \frac{\lambda}{L_PL_Q}\sum_{l_P=1}^{L_P}\sum_{l_Q=1}^{L_Q}(p^{(l_P)}p^{(l_Q)}) \end{aligned} \tag{26}$$

Thus, for $ARImp(\mathcal{M}^{(P)},\mathcal{M}^{(Q)})$, from Eq. (15), the numerator is given by

$$t_0 - t_3 = 0 \tag{27}$$

and the denominator is given by

$$0.5(t_1+t_2)-t_3 \approx 0.5\left(\frac{\lambda}{L_P}\sum_{l_P=1}^{L_P} p^{(l_P)}+\frac{\lambda}{L_Q}\sum_{l_Q=1}^{L_Q} p^{(l_Q)}\right)-\frac{\lambda}{L_P L_Q}\sum_{l_P=1}^{L_P}\sum_{l_Q=1}^{L_Q}(p^{(l_P)}p^{(l_Q)}) \tag{28}$$

In general, the denominator is not equal to zero. Since the numerator is equal to zero, we can conclude that $\mathit{ARImp}(\mathcal{M}^P,\mathcal{M}^Q)=0$. □

4.2. Computation of ARImp/ARImm in a distributed environment

Another desirable property of ARImp/ARImm is that these measures can be readily computed in a distributed environment. Specifically, we consider two main scenarios:

(i) ARImp/ARImm can be effectively used to evaluate the clustering quality when the data set is partitioned into different subsets for processing, each of which is assigned to a node in a distributed system. In real world applications, we need to handle large data sets, and the cluster ensemble technique provides an effective framework to perform this task in a distributed manner. Specifically, data might be divided into overlapping subsets, and each subset is clustered in one of the nodes in a distributed system. Finally, all the clustering solutions are aggregated to provide the consensus. However, since each clustering solution associated with each node of the system is related to only a subset of the complete data set, there are currently no clustering measures for evaluating the clustering quality in this case. It is interesting that ARImp/ARImm can be used to solve this problem since they are based on the consensus matrices rather than the individual clustering solutions themselves.

(ii) ARImp/ARImm can be effectively computed in a distributed manner. Specifically, two consensus matrices can be divided into a number of different sub-matrices, and these sub-matrices can be used to reconstruct the two consensus matrices. For example, assume that two 6 × 6 consensus matrices are divided into four 3 × 3 sub-matrices respectively, it is straightforward to observe that we can compute the ARImp/ARImm based on factors computed with these sub-matrices. Specifically, for the z-th sub-matrix of the two consensus matrices, t_0^z,t_1^z,t_2^z can be computed using Eq. (14). Thus t_0,t_1,t_2,t_3 of the ARImm on the two consensus matrices can be computed in a distributed manner as follows:

$$t_0=\sum_z t_0^z,\quad t_1=\sum_z t_1^z,\quad t_2=\sum_z t_2^z,\quad t_3=\frac{2t_1t_2}{N(N-1)} \tag{29}$$

This distributed computation approach for ARImp/ARImm could be useful in scenarios in which the issues of performance requirement and/or data security are important.

4.3. Measuring the degree of uncertainty of a consensus matrix

One of the most important properties of ARImm is its ability to measure the degree of uncertainty of a consensus matrix. We consider a simple example. Assume that for three data points x_1,x_2,x_3, there are two consensus matrices $\mathcal{M}^{(1)}=[0\ 1\ 0;\ 1\ 0\ 0;\ 0\ 0\ 0]$ and $\mathcal{M}^{(2)}=[0\ 0.5\ 0;\ 0\ 0.5\ 0;\ 0\ 0\ 0]$. $\mathcal{M}^{(1)}$ shows that x_1 and x_2 belong to the same cluster and x_3 is assigned to another cluster, all with probability 1. Similarly, $\mathcal{M}^{(2)}$ shows that x_1 and x_2 belong to the same cluster with probability 0.5 and x_3 is assigned to another cluster with probability 1. For an effective matrix similarity evaluation function $\mathit{sim}()$, we would like to argue that $\mathit{sim}(\mathcal{M}^{(1)},\mathcal{M}^{(1)}) > \mathit{sim}(\mathcal{M}^{(2)},\mathcal{M}^{(2)})$. This is because there is no uncertainty within $\mathcal{M}^{(1)}$ while uncertainty does exist within $\mathcal{M}^{(2)}$. To our best knowledge, there are no other general matrix comparison measures with this discrimination property. It is interesting that ARImm has this very property, as $\mathit{ARImm}(\mathcal{M}^{(1)},\mathcal{M}^{(1)})=1$, $\mathit{ARImm}(\mathcal{M}^{(2)},\mathcal{M}^{(2)})=0.4$, which allows the discrimination of their difference.

One of the possible applications of this property of ARImm is to evaluate the diversity of a cluster ensemble. Note that $\mathit{ARImm}(\mathcal{M},\mathcal{M})$ can be computed based on only the consensus matrix $\mathcal{M}$ instead of the individual clustering solutions, which is beyond the capability of the traditional measure Pairwise Normalized Mutual Information (PNMI). Another attractive property of $\mathit{ARImm}(\mathcal{M},\mathcal{M})$ is that we can derive the main property of its lower bound (the upper bound of either $\mathit{ARImm}(\mathcal{M},\mathcal{M})$ or PNMI is 1 when all the individual clustering solutions are identical). It is also notable that there is no previous study on the lower bound of PNMI in related works.

Proposition 4. *Lower bound of $\mathit{ARImm}(\mathcal{M},\mathcal{M})$. Given a consensus matrix $\mathcal{M}$ constructed from a set of random partitions $\{P^{(l)}\}_{l=1}^{L}$ with the maximum possible number of clusters K_{max}. We can determine that the lower bound of $\mathit{ARImm}(\mathcal{M},\mathcal{M})$ is a constant value for each fixed K_{max} and L.*

Proof. Since $\mathcal{M}_{ij}=(1/L)\sum_{l=1}^{L} M_{ij}^{(l)}$, we have

$$\begin{aligned}
t_0 &= \sum_{i\neq j}\frac{\mathcal{M}_{ij}\mathcal{M}_{ij}}{2}=\sum_{i\neq j}\frac{\frac{1}{L}\sum_{l_1=1}^{L} M_{ij}^{(l_1)}\cdot\frac{1}{L}\sum_{l_2=1}^{L} M_{ij}^{(l_2)}}{2}\\
&= \frac{1}{L}\sum_{l_1=1}^{L}\sum_{i\neq j}\frac{\frac{1}{L}(M_{ij}^{(l_1)}M_{ij}^{(l_1)}+M_{ij}^{(l_1)}\sum_{l_2\neq l_1} M_{ij}^{(l_2)})}{2}\\
&= \frac{1}{L^2}\sum_{l_1=1}^{L}\sum_{i\neq j}\frac{M_{ij}^{(l_1)}}{2}+\frac{1}{L^2}\sum_{l_1=1}^{L}\sum_{l_2\neq l_1}\frac{N(N-1)\overline{M_{ij}^{(l_1)}M_{ij}^{(l_2)}}}{2}\\
&= \frac{1}{L^2}\sum_{l_1=1}^{L}\frac{N(N-1)\overline{M_{ij}^{(l_1)}}}{2}+\frac{1}{L^2}\sum_{l_1=1}^{L}\sum_{l_2\neq l_1}\frac{N(N-1)\overline{M_{ij}^{(l_1)}M_{ij}^{(l_2)}}}{2}\\
&\approx \frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}+\frac{\lambda}{L^2}\sum_{l_1=1}^{L}\sum_{l_2\neq l_1} p^{(l_1)}p^{(l_2)}\\
&= \frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}+\frac{\lambda}{L^2}\sum_{l_1=1}^{L}\sum_{l_2=1}^{L}(p^{(l_1)}p^{(l_2)})-\frac{\lambda}{L^2}\sum_{l_1=1}^{L}(p^{(l_1)}p^{(l_1)})\\
&= \frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}+\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}\sum_{l_2=1}^{L} p^{(l_2)}-\frac{\lambda}{L^2}\sum_{l_1=1}^{L}(p^{(l_1)}p^{(l_1)})
\end{aligned} \tag{30}$$

Also, from the proof of Proposition 3, we obtain

$$t_1=t_2\approx\frac{\lambda}{L}\sum_{l=1}^{L} p^{(l)} \tag{31}$$

$$t_3=\frac{2t_1t_2}{N(N-1)}\approx\frac{1}{\lambda}\left(\frac{\lambda}{L}\sum_{l=1}^{L} p^{(l)}\right)^2=\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}\sum_{l_2=1}^{L} p^{(l_2)} \tag{32}$$

Thus, we can obtain the numerator and the denominator respectively as follows:

$$\begin{aligned}
t_0-t_3 &= \left(\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}+\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}\sum_{l_2=1}^{L} p^{(l_2)}-\frac{\lambda}{L^2}\sum_{l_1=1}^{L}(p^{(l_1)}p^{(l_1)})\right)\\
&\quad -\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}\sum_{l_2=1}^{L} p^{(l_2)}=\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}-\frac{\lambda}{L^2}\sum_{l_1=1}^{L}(p^{(l_1)}p^{(l_1)})
\end{aligned} \tag{33}$$

$$0.5(t_1+t_2)-t_3 \approx 0.5\left(\frac{\lambda}{L}\sum_{l=1}^{L} p^{(l)}+\frac{\lambda}{L}\sum_{l=1}^{L} p^{(l)}\right)-\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}\sum_{l_2=1}^{L} p^{(l_2)}$$
$$=\frac{\lambda}{L}\sum_{l=1}^{L} p^{(l)}-\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}\sum_{l_2=1}^{L} p^{(l_2)} \tag{34}$$

Therefore, we obtain

$$ARImm(\mathcal{M},\mathcal{M}) = \frac{t_0-t_3}{0.5(t_1+t_2)-t_3}$$
$$\approx \frac{\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}-\frac{\lambda}{L^2}\sum_{l_1=1}^{L}(p^{(l_1)}p^{(l_1)})}{\frac{\lambda}{L}\sum_{l=1}^{L} p^{(l)}-\frac{\lambda}{L^2}\sum_{l_1=1}^{L} p^{(l_1)}\sum_{l_2=1}^{L} p^{(l_2)}} = \frac{\alpha-\beta}{L(\alpha-\alpha^2)} \tag{35}$$

where

$$\alpha = \frac{1}{L}\sum_{l=1}^{L} p^{(l)} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{K^{(l)}} = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=2}^{K_{max}}\frac{prob(K^{(l)}=k)}{k}$$
$$=\frac{1}{L}\sum_{l=1}^{L}\sum_{k=2}^{K_{max}}\left(\frac{1}{K_{max}-1}\frac{1}{k}\right)=\frac{1}{K_{max}-1}\sum_{k=2}^{K_{max}}\frac{1}{k}=\frac{\sum_{k=1}^{K_{max}}\frac{1}{k}-1}{K_{max}-1} \tag{36}$$

and

$$\beta = \frac{1}{L}\sum_{l=1}^{L}(p^{(l)}p^{(l)}) = \frac{\sum_{k=1}^{K_{max}}\frac{1}{k^2}-1}{K_{max}-1} \tag{37}$$

It is interesting to note that the term $\sum_{k=1}^{K_{max}}(1/k)$ in Eq. (36) is the well-known finite harmonic series, and $\sum_{k=1}^{K_{max}}(1/k^2)$ in Eq. (37) is one of the well-known finite P-series (here $P=2$). The two terms can be directly computed when K_{max} is small. When K_{max} is very large, they can be computed using the corresponding approximate formulae. Thus, it is proved that the lower bound of $ARImm(\mathcal{M},\mathcal{M})$ converges to a constant value for each fixed K_{max} and L. □

5. Experiments

In this section, we conduct a number of experiments to verify the properties of our proposed measures ARImp and ARImm, with comparison to other reference measures. We also introduce several applications of ARImp and ARImm, which illustrate their usefulness in different scenarios.

5.1. Properties of ARImp and ARImm

The experiments in this subsection are conducted for the following purposes: (i) to verify the properties of ARImp and ARImm; (ii) to investigate the effect of different factors, such as the number of clusters K, the number of points N and the number of clustering solutions L. In the propositions, we adopt descriptions such as "if N is large". These experiments can provide guidelines to answer the question "how large is enough for N?"; and (iii) to confirm the validity of approximating the mean of the observed values of the random variable using its expectation for the range of data set sizes used in our experiments.

5.1.1. Detection of uncorrelatedness

We first design experiments in a similar setting as those in [29] to evaluate the performance of ARImp: consider a data set with N points and a maximum number of clusters K_{max}. For each selected number of clusters K from 2 to K_{max}, we specify a corresponding ground truth partition, and generate L independent clusterings, with their corresponding numbers of clusters selected from 2 to $2K$. For a thorough study, we have conducted a large number of trials for different choices of N and K. The following plots are shown in the first row of Fig. 2: (i) ANMI between the clustering solutions and the true labels, and (ii) ARImp between the consensus matrices of the individual clustering solutions and the true labels. From the figure, we can observe that ANMI in general increases as K increases (especially when the ratio N/K is small), while ARImp is close to zero and does not sensitively depend on K, which corresponds to the same behavior of ARI. We can also observe that the variation of ARImp becomes smoother when N is large enough (e.g., $N>500$). We also investigate the similar issue for ARImm and the other measures based on matrix comparison. Consider a data set with N points and a maximum cluster number K_{max}. For each selected true number of clusters K from 2 to K_{max}, two set of clusterings are generated. The first

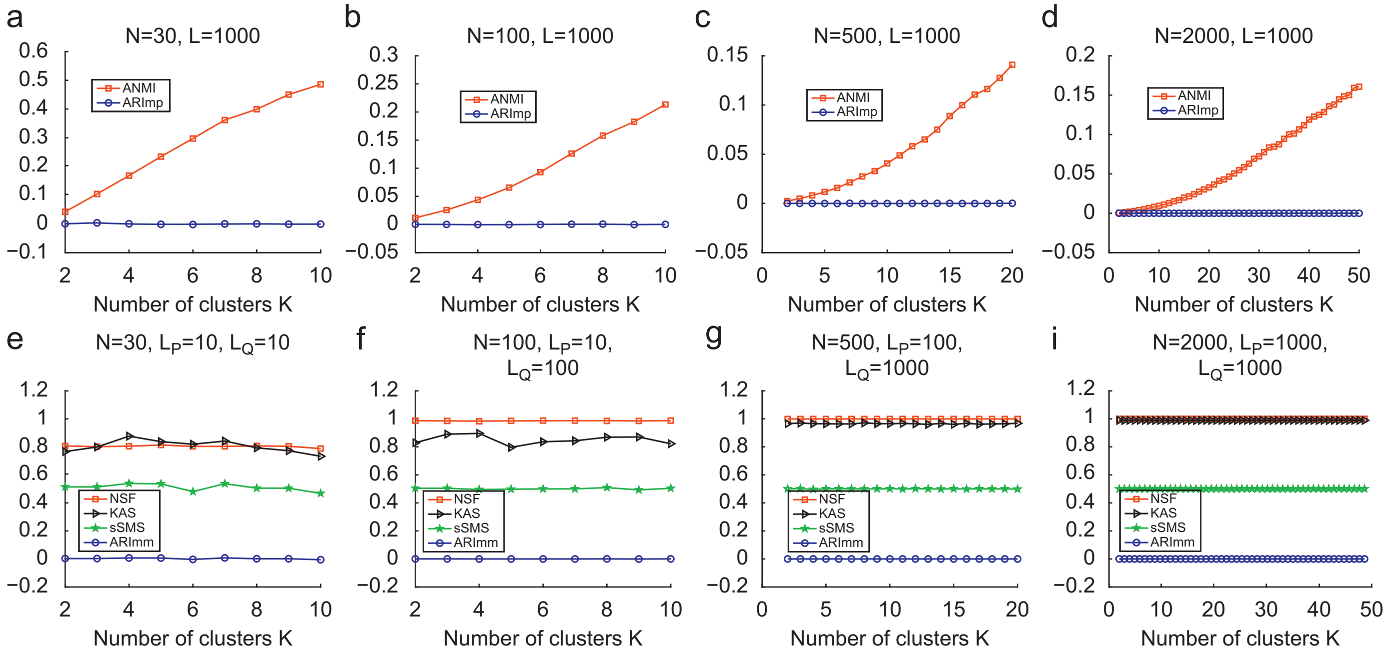

Fig. 2. Detection of uncorrelatedness. For random partitions, ARImp and ARImm are close to zero, thus inherits the same desirable property of ARI.

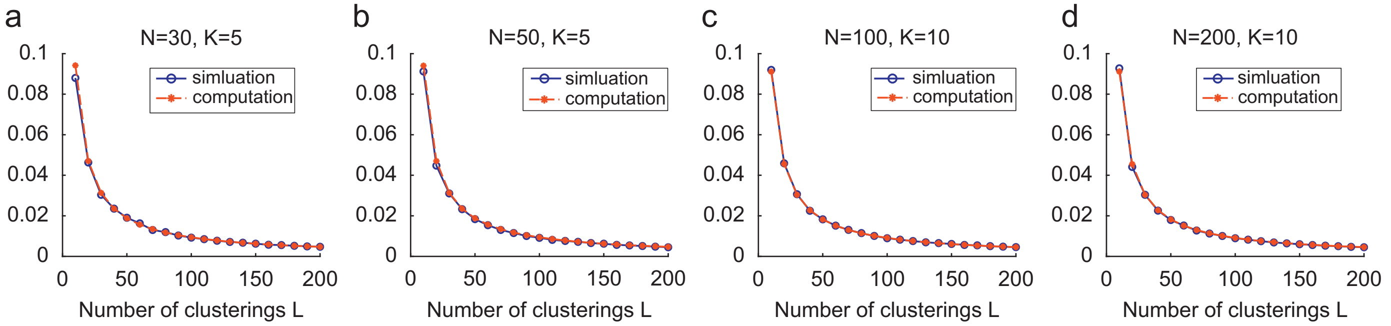

Fig. 3. Lower bound of *ARImm*($\mathcal{M}$,$\mathcal{M}$). The derived theoretical results agree with the simulation results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Summary of the UCI data sets used in the experiments.

Data set	Class	Instance	Dimension
UCI-chart	6	600	60
UCI-glass	6	214	9
UCI-iris	3	150	4
UCI-pima	2	768	8
UCI-wine	3	178	13
UCI-vehicle	4	846	18
UCI-BCW-O	2	699	9
UCI-BCW-D	2	569	30

cluster ensembles are generated from L_P and L_Q random independent clusterings, and with their cluster numbers randomly selected from 2 to 2K. Two consensus matrices are computed based on these two clustering solutions respectively. Similarity values of these four measures are shown in the second row of Fig. 2. From the results, we can observe that for different values of N,L_P,L_Q, the other three measures are all greater than 0. Specifically, NSF and KAS tend to have values near 1, while the value of sSMS tends to 0.5 with increasing N,L_P and L_Q. However, since these clusterings are all generated randomly and independently, it is reasonable to have a small similarity value as in the case of ARI. It is interesting to observe that ARImm is close to zero and insensitive to K, which corresponds to the same desirable property of ARI.

5.1.2. *Lower bound of ARImm*($\mathcal{M}$,$\mathcal{M}$)

We generate clusterings as those in the last subsection. In Fig. 3, ARImm between an ensemble and itself are shown in blue, and the computation results based on Eq. (35) are shown in red. We can observe that the two curves are similar to each other except for some slight differences when both N and L are small. This shows that the derived theoretical results agree with the simulation results.

5.2. *Applications of ARImp and ARImm*

5.2.1. *Application 1: unsupervised filtering of cluster ensemble methods (ARImp)*

We conduct experiments using several popular public data sets obtained from the well-known UCI machine learning repository,[1] which are usually used in clustering problems. All of the data sets used in the experiments are summarized in Table 2.

For each data set, 600 clustering solutions are generated using K-means with three different methods, as performed in [3]. In each case, 50 clustering solutions are sampled at random, and the final clustering solution is obtained based on the true class number K using the following different cluster ensemble methods: the cluster based Similarity Partitioning Algorithm (CSPA) [1], the hypergraph based Meta Clustering Algorithm (MCLA) [1], the Hypergraph Partitioning Algorithm (HGPA) [1], the Hybrid Bipartite Graph Formulation (HBGF) algorithm [4], Normalized cut based algorithm (NCUT) [5], and the hierarchical agglomerative clustering algorithms with Average Link (AL) [7]. The scatter plots constructed from 20 trials are shown in Fig. 4. In the plots, the horizontal axis corresponds to the ARImp value between the consensus matrix and the final clustering solution, and the vertical axis corresponds to the ARI value between the ground truth and the current clustering solution. From the figure, we can observe that, in most of the cases, approaches with associated large ARImp values and small variances in general also have large ARI values, while those with associated low ARImp values and large variances tend to have low ARI values. As a result, for a large variety of data sets, we can use ARImp to identify less effective cluster ensemble approaches associated with low ARImp values or high ARImp variances.

5.2.2. *Application 2: measuring the degree of uncertainty (ARImm)*

In this subsection, we evaluate the diversity of the clustering solutions using ARImm and PNMI. We first generate 600 clustering solutions as in the last subsection, and cluster these solutions into three clusters using spectral clustering as performed in [3]. The smallest cluster is selected as the base group, and ARImm and PNMI are computed for the clusterings in the base group. Then we randomly divide the other two clusters into four partitions. The four partitions of the second smallest cluster are added to the base group one by one, and the corresponding ARImm and PNMI are computed. Finally, the four partitions of the largest cluster are also added to the base group, and the corresponding ARImm and PNMI are also computed. In this way, we can investigate the diversity of the 600 clusterings under nine different conditions. ARImm and PNMI curves under these different conditions are shown in Fig. 5, and the correlation value of ARImm and PNMI for each data set is shown in the titles. It is interesting to observe that the ARImm curves are similar to those of PNMI and they are highly correlated (most of the correlation values are above 0.95). As a result, we can conclude that ARImm can measure the degree of uncertainty among a number of clustering solutions, as in the case of PNMI, with the availability of only the consensus

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

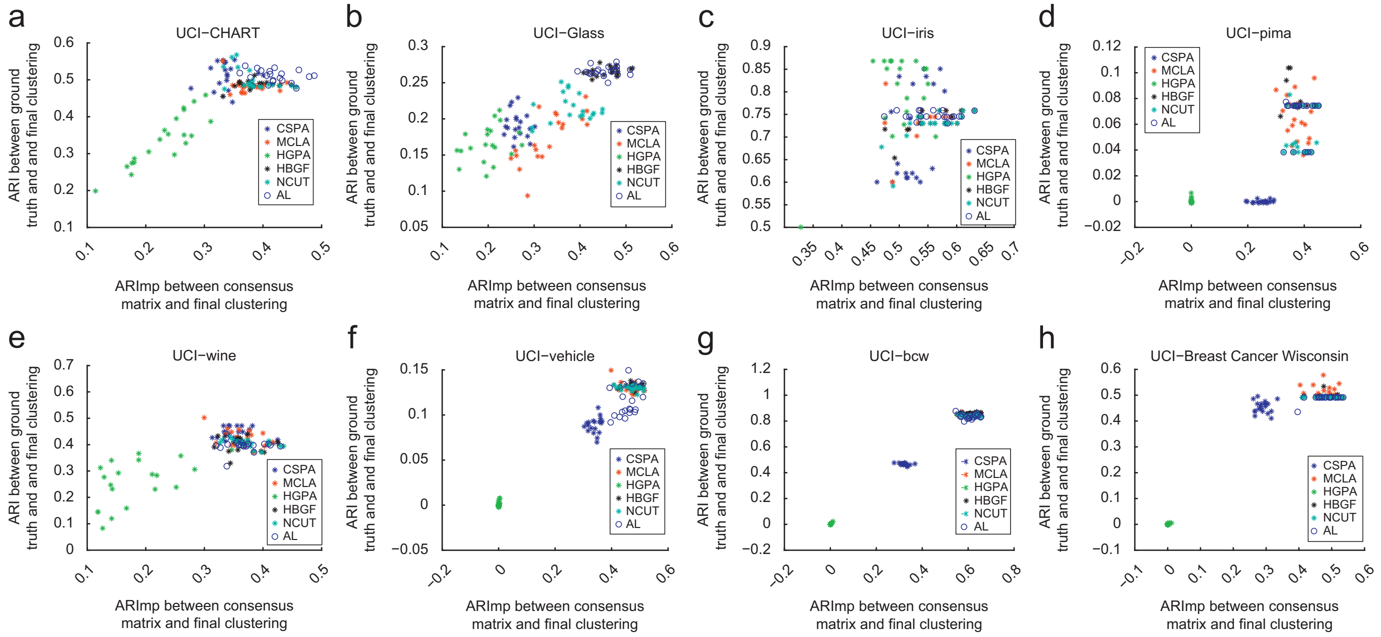

Fig. 4. Unsupervised filtering of cluster ensemble methods. Less effective cluster ensemble approaches can be identified if they have low ARImp values and large variances.

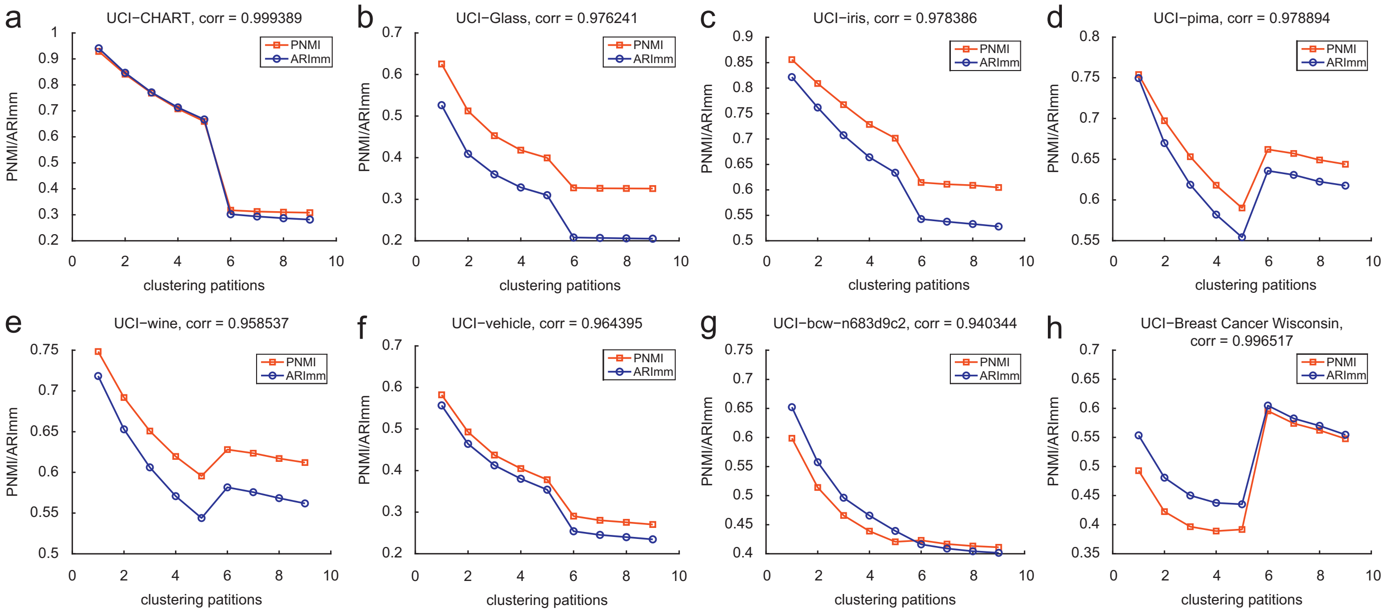

Fig. 5. Measuring diversity of cluster ensembles: PNMI vs. ARImm. Curves of PNMI and ARImm are highly correlated.

matrix, rather than requiring access to each individual clustering solution.

5.2.3. Application 3: measuring different similarity matrices (ARImp and ARImm)

As described in Introduction section, there are more application scenarios for ARImp and ARImm beyond the context of clustering ensemble. In this subsection, we introduce their capabilities to evaluate different similarity matrices. In a number of different application scenarios, different similarity matrices of pairwise relationship are usually available (we only discuss the most popular form of similarity matrices whose elements range from 0 to 1). However, there are yet no effective measures to evaluate the consistency either between different matrices with the ground truth labels or among different matrices. ARImp and ARImm could be used in these scenarios. We consider the analysis of data set using two different normalization methods: (i) min–max normalization which maps each feature component of the data into the range [0, 1] and (ii) *z*-score normalization which normalizes each feature component of the data by subtracting the mean of each component and dividing by the standard deviation.

Given two normalized points x_1 and x_2, the Gaussian kernel similarity $\mathcal{Z}(x_1,x_2)=\exp(-|x_1-x_2|^2/2)$ is used to construct the similarity matrix. To analyze the similarity matrix, a certain proportion of the entries in the matrix are replaced by the ground truth value: $\mathcal{Z}(x_1,x_2)=1$ if x_1 and x_2 belong to the same class, and $\mathcal{Z}(x_1,x_2)=0$ otherwise. ARImp curves between the group truth labels and the similarity matrices of the two kinds of normalized data are shown in Fig. 6, with the proportion of replaced entries ranging from 0.1 to 0.9. From this figure, we can observe that for all the data sets, with the proportion of replaced entries increasing, ARImp values also increase accordingly.

We also compute the similarity value between each pair of these similarity matrices using ARImm, the Normalized Similarity using the Frobenius norm (NSF), the kernel alignment similarity (KAS) and the scaled Standardized Mantel Statistic (sSMS). These results are shown in Figs. 7–10 respectively, where the darkness of the gray level denotes the degree of similarity. From the figures, we can observe that for all the data sets, the degree of darkness increases in the bottom right direction, which are in accordance with the proportion of replaced entries. Also, we can see that the diagonal elements represent the value of ARImm between each similarity matrix and itself. We also observe that

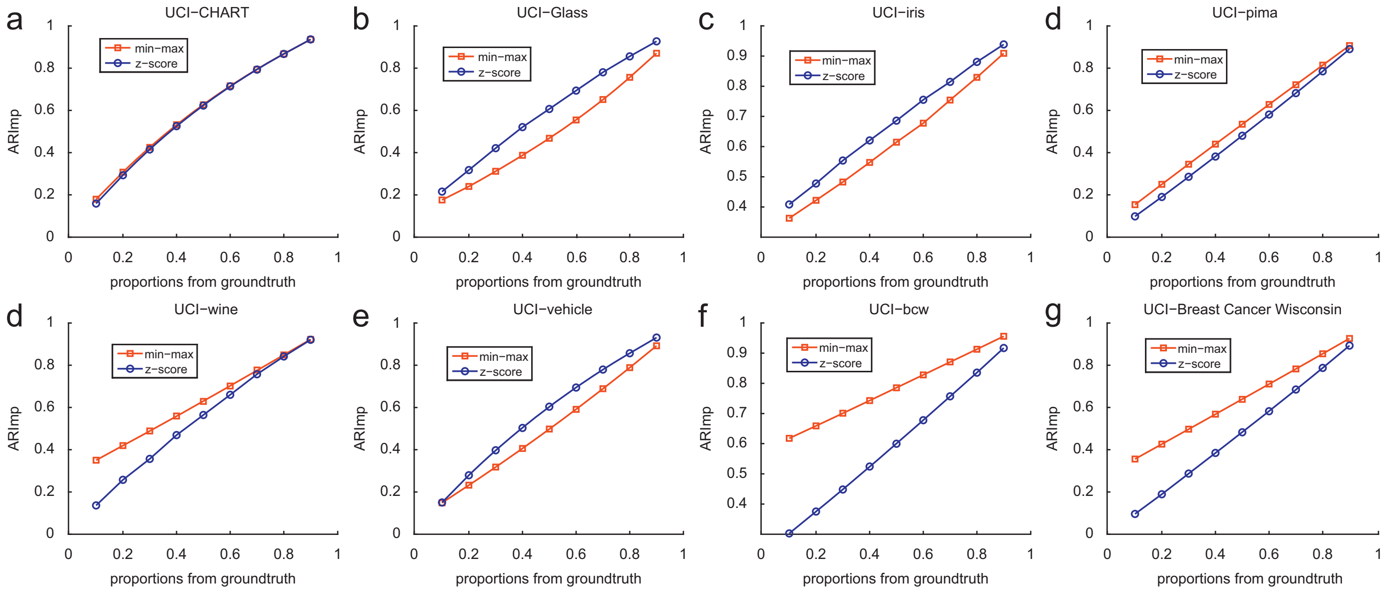

Fig. 6. Measuring similarity matrices with ground truth labels: ARImp. It can be seen that the ARImp values increase accordingly when the proportion from ground truth increases.

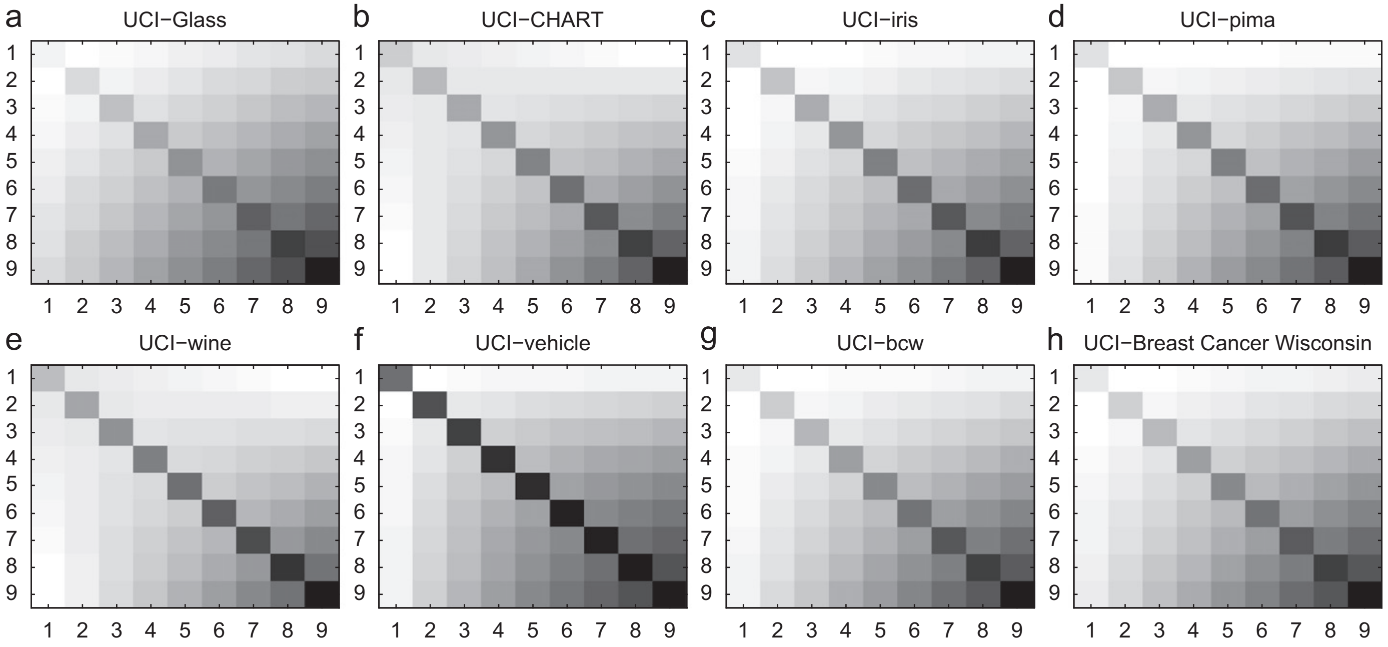

Fig. 7. Comparing the similarity between different similarity matrices: ARImm. It can be seen that the ARImm values increase accordingly when the proportion from ground truth increases.

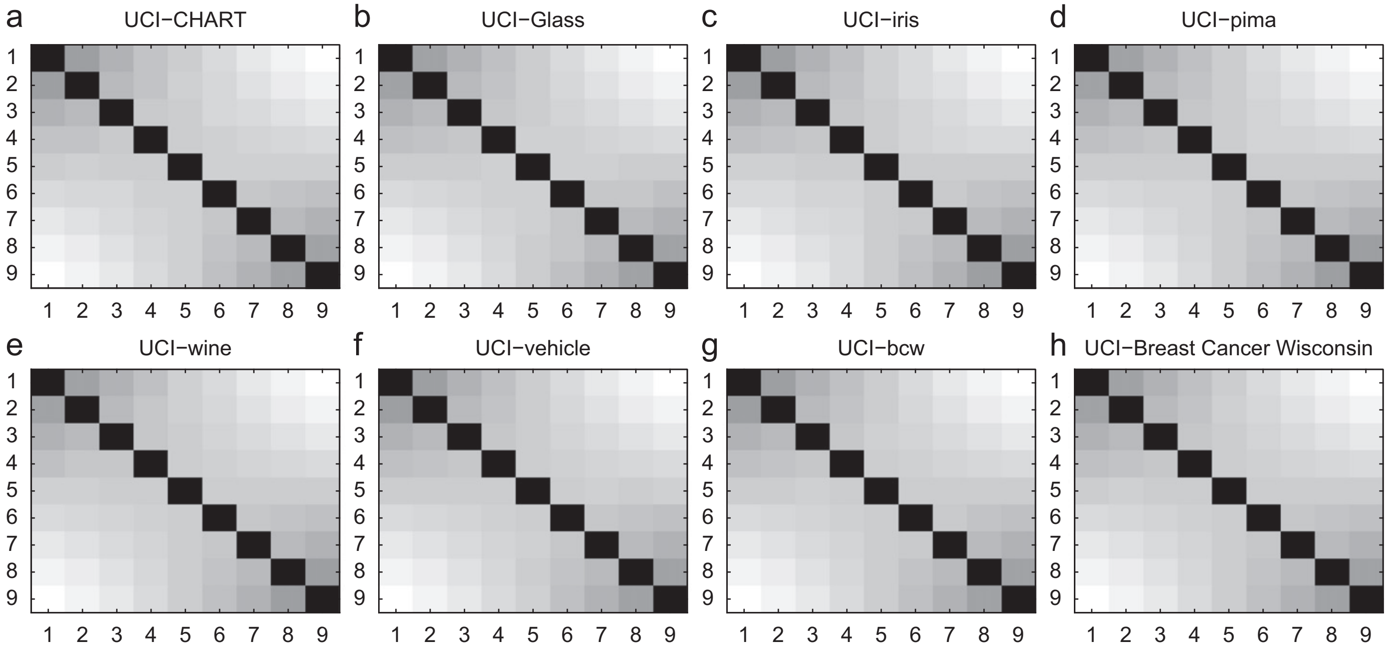

Fig. 8. Comparing the similarity between different similarity matrices: NSF.

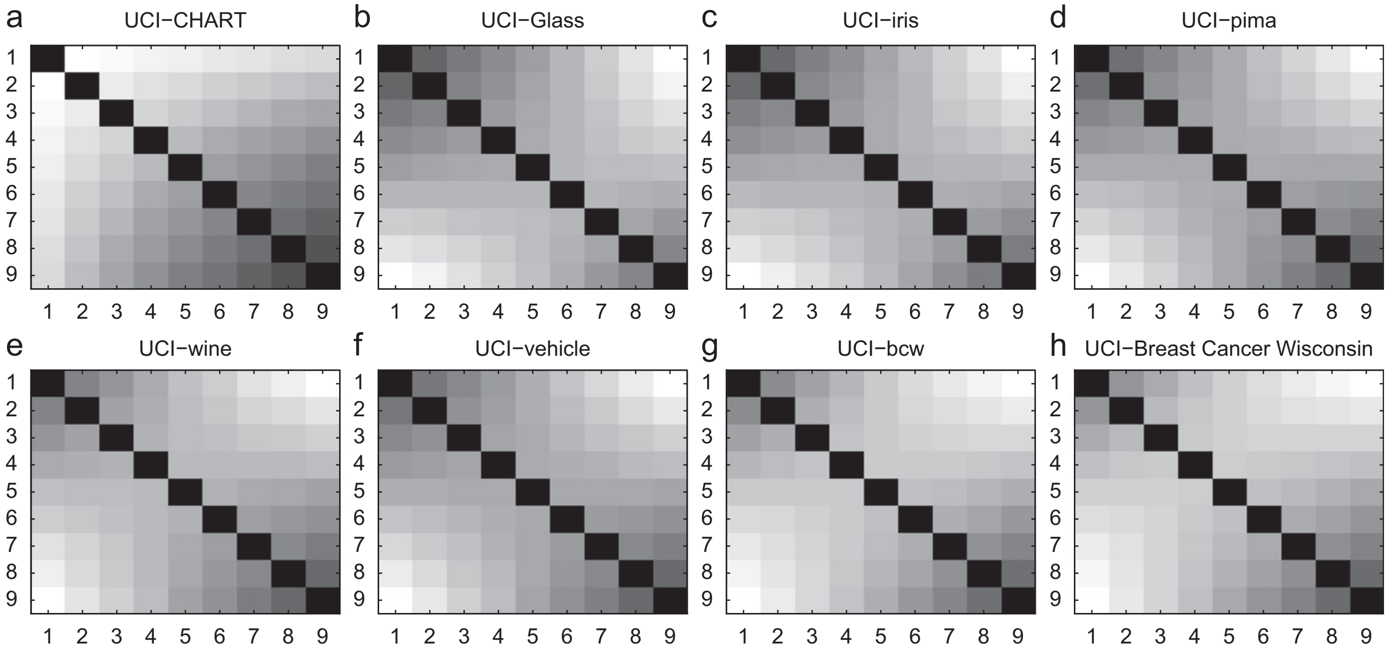

Fig. 9. Comparing the similarity between different similarity matrices: KAS.

for the measures NSF, KAS and sSMS in Figs. 8–10 respectively, the diagonal blocks are distinctively dark. On the other hand, for the measure ARImm, the transition between the gray levels is more gradual, as shown in Fig. 7. This indicates that compared to the other measures, ARImm can distinguish subtle differences between the similarity matrices associated with different degrees of uncertainty, which are consistent with the observations in Application 2. This confirms the capability of ARImm to evaluate the degree of uncertainty of the similarity matrices beyond the scope of existing algorithms.

6. Conclusions

In this paper, we generalize the popular Adjusted Rand Index (ARI) to two new measures, ARImp and ARImm, which can be used to evaluate the consistency between clustering solutions and the consensus matrix in a cluster ensemble, or between two different consensus matrices. Desirable properties of ARImp and ARImm are investigated from the perspectives of both theoretical analysis and simulation experiments. We also conduct a number of experiments on several UCI data sets to show the

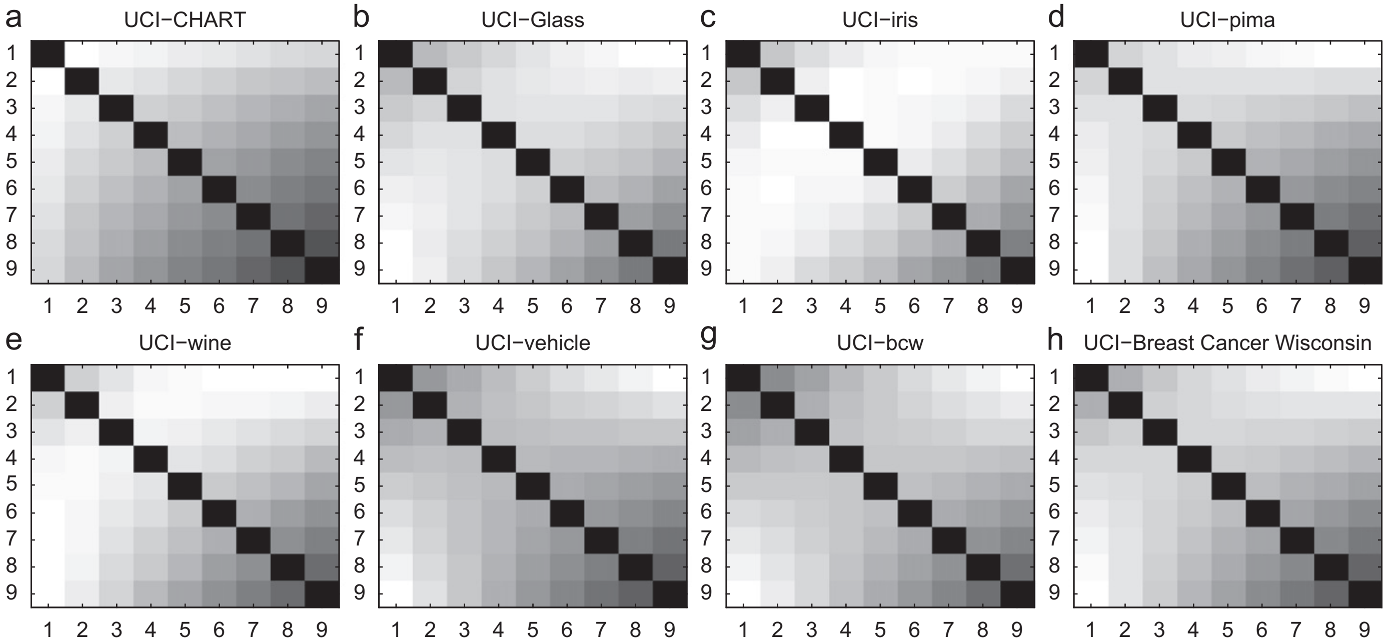

Fig. 10. Comparing the similarity between different similarity matrices: sSMS.

usefulness and effectiveness of the two proposed measures in practical applications.

Acknowledgment

The work described in this paper was supported by a grant from the City University of Hong Kong [Project No. 7008044].

References

[1] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, Journal of Machine Learning Research 3 (2002) 583–617.

[2] S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, Machine Learning 52 (1) (2003) 91–118.

[3] X. Fern, W. Lin, Cluster ensemble selection, in: Proceedings of the SIAM International Conference on Data Mining SDM, 2008, pp. 787–797.

[4] X.Z. Fern, C.E. Brodley, Solving cluster ensemble problems by bipartite graph partitioning, in: Proceedings of the Twenty-First International Conference on Machine Learning, 2004.

[5] Z. Yu, H. Wong, H. Wang, Graph-based consensus clustering for class discovery from gene expression data, Bioinformatics 23 (21) (2007) 2888.

[6] A.P. Topchy, A.K. Jain, W.F. Punch, Clustering ensembles: models of consensus and weak partitions, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (12) (2005) 1866–1881.

[7] A. Fred, A.K. Jain, Combining multiple clusterings using evidence accumulation, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (6) (2005) 835–850.

[8] L.I. Kuncheva, D. Vetrov, Evaluation of stability of *k*-means cluster ensembles with respect to random initialization, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (11) (2006) 1798–1808.

[9] H. Ayad, M.S. Kamel, Cumulative voting consensus method for partitions with variable number of clusters, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (1) (2008) 160–173.

[10] J. Azimi, X. Fern, Adaptive cluster ensemble selection, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence, 2009, pp. 992–997.

[11] A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation, ACM Transactions on Knowledge Discovery from Data (TKDD) 1 (1) (2007).

[12] L. Kuncheva, S. Hadjitodorov, L. Todorova, Experimental comparison of cluster ensemble methods, in: International Conference on Information Fusion, 2006, pp. 1–7.

[13] J.B. MacQueen, Some methods for classification and analysis of multivariate observations. in: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.

[14] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218.

[15] X. He, C. Ding, H. Zha, H. Simon, Automatic topic identification using webpage clustering, in: Proceedings of the 2001 IEEE International Conference on Data Mining, IEEE Computer Society, 2001, pp. 195–202.

[16] J. Neville, M. Adler, D. Jensen, Clustering relational data using attribute and link information, in: Proceedings of the IJCAI Text Mining and Link Analysis Workshop, Citeseer, 2003.

[17] P. Carrington, J. Scott, S. Wasserman, Models and Methods in Social Network Analysis, Cambridge University Press, 2005.

[18] P.W. Lord, R.D. Stevens, A. Brass, C.A. Goble, Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation, Bioinformatics 19 (10) (2003) 1275–1283.

[19] A. Schlicker, F. Domingues, J. Rahnenfuhrer, T. Lengauer, A new measure for functional similarity of gene products based on gene ontology, BMC Bioinformatics 7 (1) (2006) 302.

[20] X. Yin, J. Han, P. Yu, CrossClus: user-guided multi-relational clustering, Data Mining and Knowledge Discovery 15 (3) (2007) 321–348.

[21] F. Wang, C. Ding, T. Li, Integrated KL (*K*-means-Laplacian) clustering: a new clustering approach by combining attribute data and pairwise relations. in: Proceedings of the SIAM Conference on Data Mining, vol. 9, 2009, pp. 38–48.

[22] R.J.G.B. Campello, A fuzzy extension of the rand index and other related indexes for clustering and classification assessment, Pattern Recognition Letters 28 (7) (2007) 833–841.

[23] X.Z. Fern, C.E. Brodley, Random projection for high dimensional data clustering: a cluster ensemble approach, in: Proceedings of the International Conference on Machine Learning (ICML), 2003, pp. 186–193.

[24] N. Cristianini, J. Kandola, A. Elissee, On kernel target alignment. in: Advances in Neural Information Processing Systems, vol. 14, 2001.

[25] Y. Lin, T. Liu, C. Fuh, T. Sinica, Local ensemble kernel learning for object category recognition. in: Proceedings of IEEE CVPR, vol. 1, 2007, pp. 1–8.

[26] N. Mantel, A technique of disease clustering and a generalized regression approach, Cancer Research 27 (1967) 209–220.

[27] J.W. Schneider, P. Borlund, Matrix comparison, part 2: measuring the resemblance between proximity measures or ordination results by use of the mantel and procrustes statistics, Journal of the American Society for Information Science and Technology 58 (111) (2007) 1596–1609.

[28] P. Legendre, L. Legendre, Numerical Ecology, 2 ed., Elsevier, Amsterdam, 1998.

[29] N. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary?, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009.

Shaohong Zhang is a senior research associate in the Department of Computer Science, City University of Hong Kong. He received the PhD degree from Department of Computer Science, City University of Hong Kong. His research interests include machine learning, data mining, and bioinformatics.

Hau-San Wong is an Associate Professor in the Department of Computer Science at City University of Hong Kong. Prior to joining the City University of Hong Kong, he was a research associate in the School of Electrical and Information Engineering, the University of Sydney and a post-doctoral teaching fellow in the Department of Computer Science, Hong Kong Baptist University. His research interests include bioinformatics and machine learning.

Ying Shen is a PhD candidate in the Department of Computer Science at City University of Hong Kong. Her research interests include biological data mining, knowledge-based clustering, RNA 3D motif searching, and RNA tertiary structure prediction.