Journal of Machine Learning Research 16 (2015) 2949-2997 Submitted 10/12; Revised 2/15; Published 12/15

Comparing Hard and Overlapping Clusterings

Danilo Horta HORTA@ICMC.USP.BR
Ricardo J. G. B. Campello CAMPELLO@ICMC.USP.BR
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo – Campus de São Carlos
Caixa Postal 668, 13560-970, São Carlos-SP, Brazil

Editor: Marina Meila

Abstract

Similarity measures for comparing clusterings is an important component, e.g., of evaluating clustering algorithms, for consensus clustering, and for clustering stability assessment. These measures have been studied for over 40 years in the domain of exclusive hard clusterings (exhaustive and mutually exclusive object sets). In the past years, the literature has proposed measures to handle more general clusterings (e.g., fuzzy/probabilistic clusterings). This paper provides an overview of these new measures and discusses their drawbacks. We ultimately develop a corrected-for-chance measure (13AGRI) capable of comparing exclusive hard, fuzzy/probabilistic, non-exclusive hard, and possibilistic clusterings. We prove that 13AGRI and the adjusted Rand index (ARI, by Hubert and Arabie) are equivalent in the exclusive hard domain. The reported experiments show that only 13AGRI could provide both a fine-grained evaluation across clusterings with different numbers of clusters and a constant evaluation between random clusterings, showing all the four desirable properties considered here. We identified a high correlation between 13AGRI applied to fuzzy clusterings and ARI applied to hard exclusive clusterings over 14 real data sets from the UCI repository, which corroborates the validity of 13AGRI fuzzy clustering evaluation. 13AGRI also showed good results as a clustering stability statistic for solutions produced by the expectation maximization algorithm for Gaussian mixture. Implementation and supplementary figures can be found at `http://sn.im/25a9h8u`.

Keywords: overlapping, fuzzy, probabilistic, clustering evaluation

1. Introduction

Clustering is a task that aims to determine a finite set of categories (clusters) to describe a data set according to similarities/dissimilarities among its objects (Kaufman and Rousseeuw, 1990; Everitt et al., 2001). Several clustering algorithms are published every year, which makes developing of effective measures to compare clusterings indispensable (Vinh et al., 2009, 2010). Clustering algorithm A is commonly considered better than B for a given data set X if A produces clusterings that are more similar (according to a similarity measure[1] for clustering) to a reference solution for X than those produced by B. Similarity measures are also used for consensus clustering, clustering stability assessment, and even for quantifying information loss (Strehl and Ghosh, 2003; Monti et al., 2003; Yu

1. Note that a dissimilarity/distance measure can always be cast into a similarity measure. For comparison purposes, we transformed dissimilarity/distance measures into similarity measures in this work.

et al., 2007; Beringer and Hllermeier, 2007; Vinh and Epps, 2009). A consensus clustering technique aims to find a high-quality clustering solution by combining several (potentially poor) solutions obtained from different methods, algorithm initializations, or perturbations of the same data set. This combination is achieved by producing a solution that shares the most information, quantified by a similarity measure, with the original solutions (Strehl and Ghosh, 2003). In the context of clustering stability assessment, the method used to generate a set of clustering solutions is considered stable if the set shows low variation, which is considered a desirable quality (Kuncheva and Vetrov, 2006). One can apply a clustering algorithm several times to subsamples of the original data set for any numbers of clusters, producing a set of clusterings for each number of clusters. The number of clusters for which the set of solutions is less diverse is considered a good estimate of the true number of clusters (Borgelt and Kruse, 2006; Vinh and Epps, 2009). Another interesting application of similarity measures is in the quantification of information loss (Beringer and Hllermeier, 2007). To increase efficiency (e.g., in the context of data stream clustering), one can first map the data into a low-dimensional space and cluster the transformed data. If the transformation is almost lossless, the clustering structures in the two spaces should be highly similar; a similarity measure can be used to assess this.

Several established measures are suitable for comparing exclusive hard clusterings (EHCs) (Albatineh et al., 2006; Meila, 2007; Vinh et al., 2009, 2010), i.e., clusterings in which each object exclusively belongs to one cluster. Examples of popular measures are the Rand index (RI) (Rand, 1971), adjusted Rand index (ARI) (Hubert and Arabie, 1985), Jaccard index (JI) (Jaccard, 1908), mutual information (Strehl and Ghosh, 2003), and variation of information (VI) (Meila, 2005). Bcubed (BC) (Bagga and Baldwin, 1998; Amigó et al., 2009) is a measure for evaluating coreferences (e.g., a set of pronouns referring to the same noun in a paragraph) in the natural language processing field. Coreferences can also be viewed as EHCs (Cardie and Wagstaf, 1999), and BC satisfies some (frequently regarded as) desirable properties that most well-known EHC measures do not (Amigó et al., 2009). Thus, we also include BC in this work. There are other important clustering types, e.g., fuzzy/probabilistic clustering[2] (FC), non-exclusive hard clustering (NEHC), and possibilistic clustering (PC) (Campello, 2010; Anderson et al., 2010), that are not assessed using well-established measures but that would benefit from the tasks discussed above.

Various EHC measure generalizations have recently appeared in the literature (Borgelt and Kruse, 2006; Campello, 2007; Anderson et al., 2010; Campello, 2010) to fill this gap. Unfortunately, all these measures exhibit critical problems that hinder their applicability. The RI fuzzy version by Campello (2007) does not attain its maximum (i.e., 1) whenever two identical solutions are compared, which makes it difficult to convey the similarity of the compared solutions. The same issue is exhibited by other RI generalizations (Borgelt and Kruse, 2006; Ceccarelli and Maratea, 2008; Rovetta and Masulli, 2009; Brouwer, 2009; Anderson et al., 2010; Quere and Frelicot, 2011). Moreover, most of the proposed measures are not corrected for randomness, i.e., they do not provide a constant average evaluation

2. The usage of "fuzzy" or "probabilistic" depends on the interpretation of the object membership degrees given by the solution. Fuzzy c-means (Bezdek, 1981) and expectation maximization (EM) (Dempster et al., 1977) give a fuzzy and a probabilistic interpretation, respectively, although the solutions they produce come from the same domain of clusterings. We will hereafter call it fuzzy clustering in both cases for simplicity.

over sets of independently generated clusterings (constant baseline for short). In practice this means that theses measures tend to favor clusterings with certain numbers of clusters (Vinh et al., 2009, 2010), whether the compared solutions are similar or not. Additionally, several of the measures have a low sensitivity to differences in solution quality, where close evaluation values can result from comparing very similar or very different solutions.

Biclustering is also an important type of clustering solution, which is usually represented by a set of pairs $C \triangleq \{(C_1^e, C_1^c), (C_2^e, C_2^c), \ldots, (C_k^e, C_k^c)\}$. Each pair (C_r^e, C_r^c) has two non-empty sets of objects of different types. In gene expression analysis, C_r^e could be the set of genes related to the experimental conditions in C_r^c (Madeira and Oliveira, 2004). In subspace clustering, C_r^e could be the set of objects related to the object features in C_r^c (Patrikainen and Meila, 2006; Günnemann et al., 2011). We do not consider this type of clustering henceforth as it would overly extend the length and complexity of this work. Moreover, a biclustering can always be converted to an NEHC (Patrikainen and Meila, 2006), which is one of the scenarios we investigate here.

We first develop an RI generalization, called the frand index (13FRI),[3] to handle FCs. We then develop the adjusted frand index (13AFRI) by correcting 13FRI for randomness. Although the assumed randomness model is apparently unrelated to that assumed for ARI (Hubert and Arabie, 1985), we prove that 13AFRI and ARI are different formulations of the same measure in the EHC domain. Finally, we also extend the 13FRI and 13AFRI measures to the more general domain of PCs (which include the NEHC, FC, and EHC solutions as special cases, Section 3), resulting in the grand index (13GRI) and adjusted grand index (13AGRI), respectively.

We defined four clearly desirable properties that a good similarity measure should display. Under this framework, our proposed measures are empirically compared in two experiments with 32 others, out of which 28 are measures proposed in the past recent years to handle more general clusterings than EHCs. Several of the measures could not distinguish among solutions that are close to from those that are far from the reference solution according to the number of clusters in the first experiment. 13AGRI presented an evident, desirable sensitivity over the ranges of the numbers of clusters. In the second experiment, 13AGRI was the only measure that exhibited a constant baseline for all scenarios of randomly generated exclusive hard, fuzzy, non-exclusive hard, and possibilistic clusterings.

We applied 13AGRI and ARI to evaluate fuzzy c-means (Bezdek, 1981) and k-means (MacQueen, 1967) solutions, respectively, over 14 real data sets from UCI repository (Newman and Asuncion, 2010). We argue that the high correlation found between 13AGRI and ARI evaluations is an indication of the 13AGRI evaluation appropriateness for FCs. 13AGRI is also assessed as a stability statistic for FCs produced by the expectation maximization for Gaussian mixture (EMGM) (Dempster et al., 1977) algorithm.

The remainder of the paper is organized as follows. Section 2 discusses evaluation of similarity measures and establishes four desirable properties. Section 3 sets the background of the work and reviews the measures proposed in the past years to tackle more general clusterings than EHCs. Section 4 presents the 13FRI measure for handling FCs, develops a corrected-for-chance version of 13FRI named 13AFRI, and explains why 13FRI and

3. The number 13 is a reminder of the publication year of the measure (2013). We use a reminder in front of each measure acronym, except for RI, ARI, JI, and BC. This helps us identify the recently proposed measures.

13AFRI are not suitable for comparing every type of PC. Section 5 proposes the 13GRI and 13AGRI measures by addressing the issue that prevented 13FRI and 13AFRI from being appropriately applied to PCs. Section 6 deduces the asymptotic computational complexity of 13FRI, 13AFRI, 13GRI, and 13AGRI and introduces an efficient algorithm to calculate the expectations used by 13AFRI and 13AGRI. Section 7 presents four experiments, the first two to empirically evaluate the measures according to the four desirable properties. First experiment (Section 7.1) assesses how the measures behave when comparing solutions produced by clustering algorithms with reference solutions across a range of the numbers of clusters. Second experiment (Section 7.2) assesses the ability of the measures to provide unbiased evaluations in several scenarios. Third experiment (Section 7.3) compares 13AGRI and ARI evaluations of fuzzy and exclusive hard clusterings in 14 real data sets. Fourth experiment (Section 7.4) uses 13AGRI as a stability statistic for FC assessment in five real data sets. Section 8 discusses the criteria adopted to evaluate and compare the measures. Section 9 concludes the work, and Appendix proves some properties of our measures.

2. Desirable Measure Properties

Evaluating a measure for comparing clusterings is a difficult task. Partly because different applications may require different perspectives regarding the similarity between clusterings, and partly because there is no universally accepted set of properties that a measure for comparing clusterings must have. It is often the case that a measure is modified to comply with a set of desirable properties but, as a side effect, loses another set of desirable properties that it previously had. This is the case of variation of information (Meila, 2005) and its corrected-for-chance version developed in (Vinh et al., 2009, 2010), where the latter gives away the metric property to gain the property of displaying constant baseline evaluations for randomly generated solutions. There is even a result stating that no "sensible" measure for comparing clusterings will simultaneously satisfy three desirable properties (Meila, 2005).

In order to evaluate the usefulness of our proposed measure, we compare ours with the ones found in the literature over four clearly desirable properties. These properties have been chosen because they are appealing from a practical perspective and together they can unveil flaws of several existing measures according to well established intuitions. The properties are defined as follows:

- **Maximum.** A measure is told to obey this property if it attains its known maximum value whenever two equivalent solutions are compared. The maximum has to be invariant to the data set as well.
- **Discriminant.** A good measure must be able to detect the best solution among a given set of solutions.
- **Contrast.** A good measure must provide progressively better (or worse) evaluations for progressively better (or worse) solutions.
- **Baseline.** A measure that has a predetermined expected value over randomly generated solutions is told to have the baseline property (also, adjusted for chance).

It is a common practice to have the maximum equal to 1 and the baseline value equal to 0, such that having the maximum property means that the measure attains 1 when comparing

two equivalent solutions and having the baseline property means that comparing randomly generated solutions tend to give evaluations close to zero.

A measure having a known maximum that is always attained when two equivalent solutions are compared provides an objective goal (i.e., producing a clustering that attains that score) and ensures the user that a better solution can be found when the evaluation is lower than the maximum. Comparisons between evaluations of clusterings generated from different data sets may be misguided because of different extents to which variation is possible when the measure does not have a fixed maximum (Luo et al., 2009). As mentioned by Vinh et al. (2010), the fact that all of the 22 different pair counting based measures discussed in (Albatineh et al., 2006) are normalized to have a known maximum further stresses the particular interest of the clustering community in this property.

A measure may not attain its predefined maximum for the ideal solution, but still might be able to detect the best solution among a set of non-ideal solutions. This elicits the measure as having the discriminant property. This property definition naturally prompts the question "How can I know that a given solution is better than another one?" that the measure tries to answer in the first place. However, there is one situation where the answer is unquestionable: any reasonable measure should evaluate the ideal solution (i.e., the one equivalent to the reference solution) as being superior to the others. If a measure somehow evaluates a given solution better than the reference one, it is clearly flawed as a similarity measure.

We propose the contrast property because we observed in preliminary experiments that some measures would give flat evaluations over solutions progressively farther from the reference one. This behavior can be problematic when such a measure is used for assessing clustering algorithms with similar accuracy, as the measure might not be sensible enough to capture any difference.

The contrast property is also related to the useful range of a measure (Fowlkes and Mallows, 1983; Wu et al., 2009; Vinh et al., 2010). A measure can have known upper and lower bounds but its evaluations can be spread out only over a small fraction of that range in practice. As an example, for a given number of objects n, RI attains the maximum 1 for two equivalent clusterings and the minimum 0 when comparing a clustering having one cluster and a clustering having n clusters. However, it has been reported that RI provides evaluations almost always above 0.5, even when comparing randomly generated clusterings (Fowlkes and Mallows, 1983; Wu et al., 2009). Knowing beforehand the useful range (i.e., the range within which the evaluations will fall for real applications) certainly increases the intuitiveness of the measure.

The maximum property can be mathematically proved for each measure, but the other properties can only be experimentally assessed and/or disproved. The discriminant and contrast properties are somewhat subjective, but a measure that evaluates the ideal solution worse than another solution clearly does not comply with those properties. The baseline property does not specify a particular model for randomly generating solutions (and we believe that specifying one would be artificial). We thus empirically evaluate the measures regarding this property over different models of randomly generating solutions.

U/V	$\mathrm{V}_{1,:}$	$\mathrm{V}_{2,:}$	$\cdots$	$\mathrm{V}_{k_\mathrm{V},:}$	Sums
$\mathrm{U}_{1,:}$	$\mathrm{N}_{1,1}$	$\mathrm{N}_{1,2}$	$\cdots$	$\mathrm{N}_{1,k_\mathrm{V}}$	$\mathrm{N}_{1,+}$
$\mathrm{U}_{2,:}$	$\mathrm{N}_{2,1}$	$\mathrm{N}_{2,2}$	$\cdots$	$\mathrm{N}_{2,k_\mathrm{V}}$	$\mathrm{N}_{2,+}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathrm{U}_{k_\mathrm{U},:}$	$\mathrm{N}_{k_\mathrm{U},1}$	$\mathrm{N}_{k_\mathrm{U},2}$	$\cdots$	$\mathrm{N}_{k_\mathrm{U},k_\mathrm{V}}$	$\mathrm{N}_{k_\mathrm{U},+}$
Sums	$\mathrm{N}_{+,1}$	$\mathrm{N}_{+,2}$	$\cdots$	$\mathrm{N}_{+,k_\mathrm{V}}$	$\mathrm{N}_{+,+}$

Table 1: Contingency table.

3. Background and Related Work

Let $X \triangleq \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n\}$ be a data set with n objects. A clustering solution with k clusters can be represented by a matrix $\mathrm{U} \triangleq [\mathrm{U}_{r,i}] \in \mathbb{R}^{k \cdot n}$, where $\mathrm{U}_{r,i}$ expresses the membership degree of x_i to the rth cluster and U satisfies the following properties:

$$0 \leq \mathrm{U}_{r,i} \leq 1 \qquad (\forall r \in \mathbb{N}_{1,k} \text{ and } \forall i \in \mathbb{N}_{1,n}), \tag{1a}$$

$$0 < \textstyle\sum_{i=1}^{n} \mathrm{U}_{r,i} \qquad (\forall r \in \mathbb{N}_{1,k}), \text{ and} \tag{1b}$$

$$0 < \textstyle\sum_{r=1}^{k} \mathrm{U}_{r,i} \qquad (\forall i \in \mathbb{N}_{1,n}). \tag{1c}$$

We say that $\mathrm{U} \in M_\mathrm{p} \triangleq \{\mathrm{U} \in \mathbb{R}^{k \cdot n} \mid \text{satisfies Equations (1)}\}$ is a *possibilistic clustering* (PC). By adding more constraints, three other clustering types emerge: $\mathrm{U} \in M_\mathrm{f} \triangleq \{\mathrm{U} \in M_\mathrm{p} \mid \sum_{r=1}^{k} \mathrm{U}_{r,i} = 1 \quad \forall i\}$ is a *fuzzy/probabilistic clustering* (FC), $\mathrm{U} \in M_\mathrm{neh} \triangleq \{\mathrm{U} \in M_\mathrm{p} \mid \mathrm{U}_{r,i} \in \{0,1\} \quad \forall r, i\}$ is a *non-exclusive hard clustering* (NEHC), and $\mathrm{U} \in M_\mathrm{eh} \triangleq M_\mathrm{f} \cap M_\mathrm{neh}$ is an *exclusive hard clustering* (EHC) (Campello, 2010; Anderson et al., 2010). Note that $M_\mathrm{eh} \subset M_\mathrm{f}$, $M_\mathrm{eh} \subset M_\mathrm{neh}$, $M_\mathrm{f} \subset M_\mathrm{p}$, and $M_\mathrm{neh} \subset M_\mathrm{p}$ (Figure 1). Set M_p of all PCs covers the other sets, and a measure for this domain is applicable to virtually every type of clustering present in the literature.

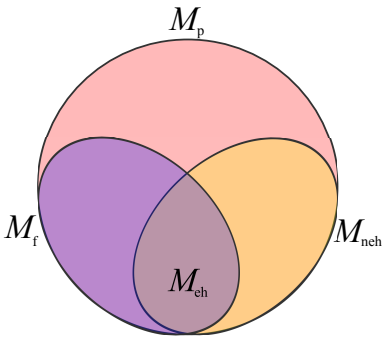


Figure 1: Venn diagram representing the relationship between clustering domains.

We believe that the most popular measures for comparing EHCs are those based on pair counting, including ARI and JI. A common approach to compute these measures begins by obtaining a contingency matrix (Albatineh et al., 2006). Let U and V be two EHCs with k_U and k_V clusters, respectively, of the same data set of n objects. Table 1 defines their contingency table, where $\mathrm{N} = \mathrm{U}\mathrm{V}^\mathrm{T}$ is the contingency matrix and $\mathrm{N}_{r,t}$ is the number

of objects that simultaneously belong to the rth cluster of U and tth cluster of V. The marginal totals $\mathrm{N}_{+,t} = \sum_{r=1}^{k_\mathrm{U}} \mathrm{N}_{r,t}$ and $\mathrm{N}_{r,+} = \sum_{t=1}^{k_\mathrm{V}} \mathrm{N}_{r,t}$ yield the cluster sizes and the grand total $\mathrm{N}_{+,+} = \sum_{r,t=1}^{k_\mathrm{U},k_\mathrm{V}} \mathrm{N}_{r,t} = n$ yields the number of objects in the data set. The contingency matrix is then used to calculate the pairing variables a (the number of object pairs in the same cluster in both U and V), b (the number of object pairs in the same cluster in U but in different clusters in V), c (the number of object pairs in different clusters in U but in the same cluster in V), and d (the number of object pairs in different clusters in both U and V) (Jain and Dubes, 1988; Albatineh et al., 2006):

$$a = \sum_{r,t=1}^{k_\mathrm{U},k_\mathrm{V}} \binom{\mathrm{N}_{r,t}}{2} = \frac{1}{2} \sum_{r,t=1}^{k_\mathrm{U},k_\mathrm{V}} \mathrm{N}_{r,t}^2 - \frac{\mathrm{N}_{+,+}}{2}, \tag{2a}$$

$$b = \sum_{r=1}^{k_\mathrm{U}} \binom{\mathrm{N}_{r,+}}{2} - a = \frac{1}{2} \sum_{r=1}^{k_\mathrm{U}} \mathrm{N}_{r,+}^2 - \frac{1}{2} \sum_{r,t=1}^{k_\mathrm{U},k_\mathrm{V}} \mathrm{N}_{r,t}^2, \tag{2b}$$

$$c = \sum_{t=1}^{k_\mathrm{V}} \binom{\mathrm{N}_{+,t}}{2} - a = \frac{1}{2} \sum_{t=1}^{k_\mathrm{V}} \mathrm{N}_{+,t}^2 - \frac{1}{2} \sum_{r,t=1}^{k_\mathrm{U},k_\mathrm{V}} \mathrm{N}_{r,t}^2, \text{ and} \tag{2c}$$

$$d = \binom{\mathrm{N}_{+,+}}{2} - (a+b+c) = \frac{1}{2}\mathrm{N}_{+,+}^2 - \frac{1}{2}\Big(\sum_{r=1}^{k_\mathrm{U}} \mathrm{N}_{r,+}^2 + \sum_{t=1}^{k_\mathrm{V}} \mathrm{N}_{+,t}^2\Big) + \frac{1}{2} \sum_{r,t=1}^{k_\mathrm{U},k_\mathrm{V}} \mathrm{N}_{r,t}^2. \tag{2d}$$

Albatineh et al. (2006) list 22 measures based on pair counting defined solely using a, b, c, and d. For example, JI and RI are respectively defined as

$$\mathrm{JI}(\mathrm{U},\mathrm{V}) \triangleq a/(a+b+c) \text{ and} \tag{3}$$

$$\mathrm{RI}(\mathrm{U},\mathrm{V}) \triangleq (a+d)/(a+b+c+d). \tag{4}$$

ARI is defined as (Hubert and Arabie, 1985)[4]

$$\mathrm{ARI}(\mathrm{U},\mathrm{V}) \triangleq \frac{a - \frac{(a+c)(a+b)}{a+b+c+d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{a+b+c+d}}. \tag{5}$$

As an alternative to the contingency matrix, one can define the pairing variables by employing the co-association matrices $\mathrm{J}^\mathrm{U} \triangleq \mathrm{U}^\mathrm{T}\mathrm{U}$ and $\mathrm{J}^\mathrm{V} \triangleq \mathrm{V}^\mathrm{T}\mathrm{V}$ (Zhang et al., 2012). When U and V are EHCs, the above definition amounts to

$$\mathrm{J}_{i,j}^\mathrm{U} = \begin{cases} 1 & \text{if } \exists r \text{ such that } \mathrm{U}_{r,i} = 1 \text{ and } \mathrm{U}_{r,j} = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

The pairing variables can be rewritten as[5]

$$\begin{array}{ll} a = \sum_{i<j} \mathrm{J}_{i,j}^\mathrm{U}\mathrm{J}_{i,j}^\mathrm{V}, & b = \sum_{i<j} \mathrm{J}_{i,j}^\mathrm{U}(1-\mathrm{J}_{i,j}^\mathrm{V}), \\ c = \sum_{i<j}(1-\mathrm{J}_{i,j}^\mathrm{U})\mathrm{J}_{i,j}^\mathrm{V}, \text{ and} & d = \sum_{i<j}(1-\mathrm{J}_{i,j}^\mathrm{U})(1-\mathrm{J}_{i,j}^\mathrm{V}). \end{array} \tag{7}$$

4. Equation (5) in (Hubert and Arabie, 1985) for ARI is defined by combinations. However, it is equivalent to Equation (5) defined here, as $a = \sum_{r,t=1}^{k_\mathrm{U},k_\mathrm{V}} \binom{\mathrm{N}_{r,t}}{2}$, $a+b = \sum_{r=1}^{k_\mathrm{U}} \binom{\mathrm{N}_{r,+}}{2}$, and $a+c = \sum_{t=1}^{k_\mathrm{V}} \binom{\mathrm{N}_{+,t}}{2}$.

5. $\sum_{i<j}$ is a shorthand for $\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}$.

BC is based on bcubed precision (BCP) and bcubed recall (BCR) (Amigó et al., 2009):

$$\mathrm{BCP}(\mathrm{U},\mathrm{V}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{n}\mathrm{J}^{\mathrm{U}}_{i,j}\mathrm{J}^{\mathrm{V}}_{i,j}}{\sum_{j=1}^{n}\mathrm{J}^{\mathrm{U}}_{i,j}} \quad \text{and} \tag{8a}$$

$$\mathrm{BCR}(\mathrm{U},\mathrm{V}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{n}\mathrm{J}^{\mathrm{U}}_{i,j}\mathrm{J}^{\mathrm{V}}_{i,j}}{\sum_{j=1}^{n}\mathrm{J}^{\mathrm{V}}_{i,j}}. \tag{8b}$$

BC is defined by default as:

$$\mathrm{BC}(\mathrm{U},\mathrm{V}) \triangleq 2\cdot\frac{\mathrm{BCP}(\mathrm{U},\mathrm{V})\cdot\mathrm{BCR}(\mathrm{U},\mathrm{V})}{\mathrm{BCP}(\mathrm{U},\mathrm{V})+\mathrm{BCR}(\mathrm{U},\mathrm{V})}.$$

3.1 Similarity Measures for Clustering

Table 2 provides an overview of recently proposed measures designed to handle more general solutions than EHCs. For each measure, this table shows the clustering types for which it was designed and the approach used in its formulation.

03VI, 03MI, and 05MI are three measures based on information theory (Mackay, 2003). Let U and V be two FCs with k_{U} and k_{V} clusters, respectively. The joint probability $\mathrm{P}(r,t)$ of an object belonging to both the rth cluster in U and tth cluster in V is defined by dividing the contingency matrix N by n, i.e. $\mathrm{P}(r,t) \triangleq \mathrm{N}_{r,t}/n$. The mutual information between U and V is defined as:

$$\mathrm{I}(\mathrm{U},\mathrm{V}) \triangleq \sum_{r,t=1}^{k_{\mathrm{U}},k_{\mathrm{V}}}\mathrm{P}(r,t)\log\left(\frac{\mathrm{P}(r,t)}{\mathrm{P}(r,\cdot)\mathrm{P}(\cdot,t)}\right),$$

where $\mathrm{P}(r,\cdot) \triangleq \sum_{t=1}^{k_{\mathrm{V}}}\mathrm{P}(r,t)$ and $\mathrm{P}(\cdot,t) \triangleq \sum_{r=1}^{k_{\mathrm{U}}}\mathrm{P}(r,t)$ are the marginals. The entropy associated with U is

$$\mathrm{H}(\mathrm{U}) \triangleq \sum_{r=1}^{k_{\mathrm{U}}}\mathrm{P}(r,\cdot)\log\left(\mathrm{P}(r,\cdot)\right).$$

The 03VI, 03MI, and 05MI measures are defined as:

$$\begin{aligned}
\mathrm{03VI}(\mathrm{U},\mathrm{V}) &\triangleq \mathrm{H}(\mathrm{U})+\mathrm{H}(\mathrm{V})-2\mathrm{I}(\mathrm{U},\mathrm{V}),\\
\mathrm{03MI}(\mathrm{U},\mathrm{V}) &\triangleq \mathrm{I}(\mathrm{U},\mathrm{V})/\sqrt{\mathrm{H}(\mathrm{U})\mathrm{H}(\mathrm{V})}, \text{ and}\\
\mathrm{05MI}(\mathrm{U},\mathrm{V}) &\triangleq 2\mathrm{I}(\mathrm{U},\mathrm{V})/(\mathrm{H}(\mathrm{U})+\mathrm{H}(\mathrm{V})).
\end{aligned}$$

We assume base two for $\log(\cdot)$ in the experiments (Section 7).

07CRI was developed based on a set-theoretic formulation of pairing variables. Let U and V be two EHCs. Let R be the set of unordered object pairs belonging to the same cluster in U, and let T be the set of unordered object pairs belonging to the same cluster in V. The usual cardinality $|R\cap T|$ yields the pairing variable a; using the same approach, variables b, c, and d can be defined by their sets. Fuzzy versions of the pairing variables were then defined by replacing the usual set operations with counterparts from fuzzy set

Measure	EHC	FC	NEHC	PC	Based on
03VI (Meila, 2003) 03MI (Strehl and Ghosh, 2003) 05MI (Fred and Jain, 2005)	*	*			Information theory
07CRI (Campello, 2007) 07CARI	*	*	*	*	Fuzzy sets (a, b, c, d)
08BRIp (Borgelt, 2007) 08BRIm	*	*			J^{U} (a, b, c, d)
09EBC (Amigó et al., 2009)	*		*		Precision/Recall
09CRI (Ceccarelli and Maratea, 2009) 09CARI	*	*	*	*	$\dot{\mathrm{N}}$ (a, b, c, d)†
09HI (Hullermeier and Rifqi, 2009)	*	*	*	*	Dist. ($\mathrm{U}_{:,i}$ and $\mathrm{U}_{:,j}$)
09RI (Rovetta and Masulli, 2009)	*	*			J^{U} (ad hoc)
09BRI (Brouwer, 2009) 09BARI	*	*	*	*	J^{U} (ad hoc)
10QRIp (Quere et al., 2010) 10QRIm	*	*	*	*	J^{U} (a, b, c, d)
10ARI (Anderson et al., 2010) 10AARI 10ARIn 10AARIn	*	*	*	*	N (a, b, c, d)*
10CSI (Campello, 2010)	*		*		ad hoc
10CF (Campello, 2010) 10CFn	*	*	*	*	Edit distance
11ARInm (Anderson et al., 2011) 11AARInm	*	*	*	*	N (a, b, c, d)*
11MD (Wang, 2010)	*		*		J^{U} (ad hoc)
11D2 (Wang, 2010)	*		*		Hamming distance
12DB (Wang, 2012)	*		*		Information theory

† The contingency matrix N used is not the same as the original one. Ceccarelli and Maratea (2009) it defined as $\dot{\mathrm{N}}_{r,t} \triangleq \sum_{i=1}^{n}(\mathrm{U}_{r,i} + \mathrm{V}_{t,i})^{\alpha}$. We adopt $\alpha \triangleq 1$ for simplicity.

* Measures 10ARIn, 10AARIn, 11ARInm, and 11AARInm use a normalized contingency matrix $\hat{\mathrm{N}}$.

Table 2: General similarity measures.

theory (Campello, 2007). Plugging the new versions of a, b, c, and d into Equations (4) and (5) resulted in 07CRI and 07CARI, respectively, where U and V are PCs.

08BRIp and 08BRIm are RI generalizations based on the definitions of a, b, c, and d given by Equations (7), where an arbitrary t-norm (from fuzzy set theory) replaces the multiplication operator used to compute $\mathrm{J}^{\mathrm{U}} = \mathrm{U}^{\mathrm{T}}\mathrm{U}$, $\mathrm{J}^{\mathrm{V}} = \mathrm{V}^{\mathrm{T}}\mathrm{V}$, and variables a, b, c, and d. We adopted the well-known product t-norm ($\top_{\mathrm{prod}}(x, y) \triangleq xy$) and minimum t-norm ($\top_{\min}(x, y) \triangleq \min\{x, y\}$) to define 08BRIp and 08BRIm, respectively.

09EBC is based on the redefinitions of BCP and BCR (Equations 8):

$$\mathrm{EBCP}(\mathrm{U}, \mathrm{V}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{n}\min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}}{\sum_{j=1}^{n}\mathrm{J}^{\mathrm{U}}_{i,j}} \text{ and} \tag{9a}$$

$$\mathrm{EBCR}(\mathrm{U}, \mathrm{V}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{n}\min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}}{\sum_{j=1}^{n}\mathrm{J}^{\mathrm{V}}_{i,j}}. \tag{9b}$$

Equations (8) and (9) are equivalent when U and V are EHCs. 09EBC is defined by default as:

$$09\mathrm{EBC}(\mathrm{U}, \mathrm{V}) \triangleq 2 \cdot \frac{\mathrm{EBCP}(\mathrm{U}, \mathrm{V}) \cdot \mathrm{EBCR}(\mathrm{U}, \mathrm{V})}{\mathrm{EBCP}(\mathrm{U}, \mathrm{V}) + \mathrm{EBCR}(\mathrm{U}, \mathrm{V})},$$

for NEHCs U and V.

09CRI and 09CARI are based on a reformulation of contingency matrix N, where the sum operator replaces the multiplication operator (i.e., $\dot{\mathrm{N}}_{r,t} \triangleq \sum_{i=1}^{n}(\mathrm{U}_{r,i} + \mathrm{V}_{t,i})$), and the subsequent pairing variable calculation uses an equivalent formulation (in the EHC domain) to that in Equations (2) (Equations (14), (15), (16), and (21) in (Ceccarelli and Maratea, 2009)). 09CRI and 09CARI are obtained by plugging these new pairing variables into Equations (4) and (5), respectively.

09HI is based on similarity calculations between the columns of U and V. Let $\mathrm{R}^{\mathrm{U}}_{i,j} \triangleq 1 - \|\mathrm{U}_{:,i} - \mathrm{U}_{:,j}\|$ and $\mathrm{R}^{\mathrm{V}}_{i,j} \triangleq 1 - \|\mathrm{V}_{:,i} - \mathrm{V}_{:,j}\|$ for all i, j be the similarities between the columns of U and V, where $\|\cdot\|$ is a norm that yields values in $[0, 1]$.[6] The degree of concordance between the distances from U and V defines the measure: $09\mathrm{HI}(\mathrm{U}, \mathrm{V}) \triangleq 1 - \sum_{i<j}|\mathrm{R}^{\mathrm{U}}_{i,j} - \mathrm{R}^{\mathrm{V}}_{i,j}|/(n(n-1)/2)$.

09RI is based on the co-association matrices J^{U} and J^{V}. The 09RI formulation given in Equation (7) of (Rovetta and Masulli, 2009) is incorrect, and Rovetta, S. kindly provided the correct formulation by personal communication, which we repeat here. Given $\mathrm{J}^{\mathrm{U}} = \mathrm{U}^{\mathrm{T}}\mathrm{U}$ and $\mathrm{J}^{\mathrm{V}} = \mathrm{V}^{\mathrm{T}}\mathrm{V}$, the following variables are computed: $\pi \triangleq \sum_{i<j}\mathrm{J}^{\mathrm{U}}_{i,j}\mathrm{J}^{\mathrm{V}}_{i,j}$, $\sigma_{\mathrm{U}} \triangleq \sum_{i<j}\mathrm{J}^{\mathrm{U}}_{i,j}$, and $\sigma_{\mathrm{V}} \triangleq \sum_{i<j}\mathrm{J}^{\mathrm{V}}_{i,j}$. The 09RI measure is then given by $1 + (2\pi - \sigma_{\mathrm{U}} - \sigma_{\mathrm{V}})/\binom{n}{2}$.

09BRI and 09BARI are based on the pairing variables defined in Equations (7). For example, variable a was defined as $(\sum_{i,j=1}^{n}\dot{\mathrm{J}}^{\mathrm{U}}_{i,j}\dot{\mathrm{J}}^{\mathrm{V}}_{i,j} - n)/2$, where the co-association matrices used are normalized: $\dot{\mathrm{J}}^{\mathrm{U}}_{i,j} \triangleq \sum_{r=1}^{k_{\mathrm{U}}}(\mathrm{U}_{r,i}\mathrm{U}_{r,j})/(\|\mathrm{U}_{:,i}\|_{\mathrm{e}}\|\mathrm{U}_{:,j}\|_{\mathrm{e}})$.[7] Plugging these new variables into Equations (4) and (5) yields 09BRI and 09BARI, respectively.

10QRIp and 10QRIm are derived from 08BRIp and 08BRIm, respectively, by normalizing J^{U} and J^{V} such that all diagonal terms equal 1, and letting U and V be PCs. The

6. We adopted the usual Euclidean norm in the experiments.
7. $\|\cdot\|_{\mathrm{e}}$ is the usual Euclidean norm.

rationale behind this normalization is that a diagonal term $\mathrm{J}^{\mathrm{U}}_{i,i}$ should always provide the maximum, as it somehow represents the degree to which object x_i is in the same cluster as itself.

The 10ARI and 10AARI pairing variables are defined using the original formulation $\mathrm{N} = \mathrm{U}\mathrm{V}^{\mathrm{T}}$ and Equations (2). Equations (4) and (5) are then applied to yield 10ARI and 10AARI, respectively. Anderson et al. (2010) noticed that at least 10ARI does not provide evaluations confined in the interval $[0, 1]$ (as RI does) for general PCs. They thus proposed the use of a normalized contingency matrix $\hat{\mathrm{N}} \triangleq \big(n/\mathrm{N}_{+,+}\big)\mathrm{N}$ to have $\hat{\mathrm{N}}_{+,+} = n$ to alleviate the above issue. We denote the normalized versions of 10ARI and 10AARI by 10ARIn and 10AARIn, respectively.

It has been observed that 10ARIn and 10AARIn do not attain their maxima whenever two equivalent solutions[8] are compared (Anderson et al., 2011). 11ARInm and 11AARInm were then defined to address this issue as:

$$\mathrm{11ARInm}(\mathrm{U}, \mathrm{V}) \triangleq \mathrm{10ARIn}(\mathrm{U}, \mathrm{V}) / \max\{\mathrm{10ARIn}(\mathrm{U}, \mathrm{U}), \mathrm{10ARIn}(\mathrm{V}, \mathrm{V})\} \text{ and}$$
$$\mathrm{11AARInm}(\mathrm{U}, \mathrm{V}) \triangleq \mathrm{10AARIn}(\mathrm{U}, \mathrm{V}) / \max\{\mathrm{10AARIn}(\mathrm{U}, \mathrm{U}), \mathrm{10AARIn}(\mathrm{V}, \mathrm{V})\}.$$

The 10CSI measure was designed to handle non-exclusive and exclusive hard clusterings. Let $\mathrm{J}^{\mathrm{U}} = \mathrm{U}^{\mathrm{T}}\mathrm{U}$ and $\mathrm{J}^{\mathrm{V}} = \mathrm{V}^{\mathrm{T}}\mathrm{V}$ be the co-association matrices, and let $\mathrm{U}_{+,i}$ and $\mathrm{V}_{+,i}$ be the number of clusters to which object x_i belongs, according to the respective solutions. The agreement and disagreement between U and V according to the relative placement of objects x_i and x_j are defined by 10CSI as:

$$a^g_{i,j} \triangleq \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\} + \min\{\mathrm{U}_{+,i}, \mathrm{V}_{+,i}\} + \min\{\mathrm{U}_{+,j}, \mathrm{V}_{+,j}\} - 2 \text{ and}$$
$$d^g_{i,j} \triangleq |\mathrm{J}^{\mathrm{U}}_{i,j} - \mathrm{J}^{\mathrm{V}}_{i,j}| + |\mathrm{U}_{+,i} - \mathrm{V}_{+,i}| + |\mathrm{U}_{+,j} - \mathrm{V}_{+,j}|$$

10CSI is given by $\sum_{i<j} a^g_{i,j} / \sum_{i<j}(a^g_{i,j} + d^g_{i,j})$, which reduces to JI in the EHC domain.

The 10CF and 10CFn measures largely differ from the others because they are not pair-based nor based on information theory. 10CF and 10CFn are somehow related to the edit distance commonly used to define the compatibility degree between two strings of text (Levenshtein, 1966). Campello (2010) defined the fuzzy transfer distance $\mathrm{F}_{\mathrm{TD}}(\mathrm{U}, \mathrm{V})$ between two PCs U and V as the minimum amount of membership degrees that must be given to and/or removed from the objects of U (V) to make this clustering equivalent to V (U). We define here 10CF as $\mathrm{10CF}(\mathrm{U}, \mathrm{V}) \triangleq 1 - \mathrm{F}_{\mathrm{TD}}(\mathrm{U}, \mathrm{V})$ such that it yields values in the interval $(-\infty, 1]$ and attains 1 iff U and V are equivalent clusterings (Campello, 2010). 10CFn is 1 minus the normalized version of F_{TD}: $\mathrm{10CFn}(\mathrm{U}, \mathrm{V}) \triangleq 1 - \mathrm{F}_{\mathrm{TD}}(\mathrm{U}, \mathrm{V})/(n \max\{k_{\mathrm{U}}, k_{\mathrm{V}}\})$. $\mathrm{10CFn}(\mathrm{U}, \mathrm{V})$ lies in the interval $[0, 1]$ (Campello, 2010).

Let U and V be two NEHCs with k_{U} and k_{V} clusters, respectively. The 11MD and 11D2 measures are defined as:

$$\mathrm{11MD}(\mathrm{U}, \mathrm{V}) \triangleq 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j=1}^{n} |\mathrm{J}^{\mathrm{U}}_{i,j} - \mathrm{J}^{\mathrm{V}}_{i,j}|}{\sum_{j=1}^{n} \max\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}} \text{ and}$$
$$\mathrm{11D2}(\mathrm{U}, \mathrm{V}) \triangleq 1 - \frac{1}{n^2}\sum_{i,j=1}^{n} |\mathrm{J}^{\mathrm{U}}_{i,j} - \mathrm{J}^{\mathrm{V}}_{i,j}|,$$

8. Clusterings U and V are equivalent iff (i) they have the same number of clusters and (ii) V can always be transformed into U by row permutations.

where $\sum_{i,j=1}^{n} |\mathrm{J}^{\mathrm{U}}_{i,j} - \mathrm{J}^{\mathrm{V}}_{i,j}|$ is the Hamming distance when U and V are EHCs.

Let A^{U} be the adjacency matrix of U defined as:

$$\mathrm{A}^{\mathrm{U}}_{i,j} \triangleq \begin{cases} 1 & \exists r : \mathrm{U}_{r,i}\mathrm{U}_{r,j} = 1 \\ 0 & \text{otherwise} \end{cases}.$$

The normalized disconnectivity of U is given by default as (Wang, 2012):

$$\mathrm{NDisc}(\mathrm{U}) \triangleq 2(1 - \frac{1}{n^2}\sum_{i,j=1}^{n} \mathrm{A}^{\mathrm{U}}_{i,j}).$$

Let R be a (possibly degenerate) clustering resulting from the intersection between the clusters of NEHCs U and V:

$$\mathrm{R}_{(r+(t-1)*k_{\mathrm{U}}),i} \triangleq \mathrm{U}_{r,i}\mathrm{V}_{t,i}.$$

The 12DB measure is defined by default as:

$$12\mathrm{DB}(\mathrm{U},\mathrm{V}) \triangleq 2 \cdot \mathrm{NDisc}(\mathrm{R}) - \mathrm{NDisc}(\mathrm{U}) - \mathrm{NDisc}(\mathrm{V}).$$

3.2 Discussion

Some authors extended pair-based measures by simply letting U and V be representations of other clustering types (i.e., others than EHC types) in the definition of contingency matrix N (e.g., Ceccarelli and Maratea, 2009; Anderson et al., 2010) or co-association matrices J^{U} and J^{V} (e.g., Borgelt and Kruse, 2006; Borgelt, 2007; Quere and Frelicot, 2011), and computing a, b, c, and d based on Equations (2) or Equations (7). However, the pairing variable equations were deduced by assuming that U and V are EHCs. Without a more principled explanation, we believe there is no reason to expect that using the same definitions would grant meaningful values to a, b, c, and d in more general circumstances. Consider the following identical EHCs:

$$\mathrm{U} \triangleq \mathrm{V} \triangleq \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}. \tag{10}$$

We have $a = 0$, $b = 0$, $c = 0$, and $d = 1$, according to the definitions given by Equations (2) and Equations (7). There is only one pair of objects, and the objects are not clustered together in both solutions. Now let

$$\dot{\mathrm{V}} \triangleq \begin{pmatrix} 0.9 & 0.0 \\ 0.1 & 1.0 \end{pmatrix} \tag{11}$$

be an FC very similar to V. Comparing U and $\dot{\mathrm{V}}$, we now have $a = -0.09$, $b = 0.1$, $c = 0.09$, and $d = 0.9$, according to Equations (2), and $a = 0$, $b = 0$, $c = -0.1$, and $d = 0.9$, according to Equations (7). It is hard to assign a meaningful interpretation when a pairing variable yields a negative value. Moreover, the obtained values are no longer equivalent to each other. This result shows that the application of Equations (2) and (7) in more general settings must indeed be accompanied by a good justification.

None of the measures 03VI, 03MI, 05MI, 07CRI, 07CARI, 08BRIp, 08BRIm, 09CRI, 09CARI, 09RI, 09BRI, 09BARI, 10QRIp, 10QRIm, 10ARI, 10AARI, 10ARIn, 10AARIn, 10CF, 11ARInm, and 12DB attain their maxima 1 whenever two equivalent solutions are compared, as Section 7.1 shows. This makes interpreting the evaluation provided by these measures difficult. Moreover, there is no reason to expect that ARI generalizations (i.e., 07CARI, 09CARI, 09BARI, 10AARI, 10AARIn, and 11AARInm) are corrected for randomness in others than in EHC scenarios simply because the original ARI has this property for EHCs (this belief is confirmed in the experiments in Section 7.2). The formulations upon which these generalized measures are based were deduced by assuming that the compared solutions are EHCs.

4. Frand Index

Given two FCs U (with k_{U} clusters) and V (with k_{V} clusters) of n objects, 13FRI recasts each into two n-by-n matrices to retain only the essential information and to facilitate the comparison. Let $\mathrm{I}_{k_{\mathrm{U}}}$ be the k_{U}-by-k_{U} identity matrix and $\mathbb{1}_{k_{\mathrm{U}}}$ be the k_{U}-by-k_{U} matrix with 1 in each entry. Define the matrices

$$\mathrm{J}^{\mathrm{U}} \triangleq \mathrm{U}^{\mathrm{T}}\mathrm{U} \quad \text{and} \tag{12a}$$

$$\mathrm{S}^{\mathrm{U}} \triangleq \mathrm{U}^{\mathrm{T}}(\mathbb{1}_{k_{\mathrm{U}}} - \mathrm{I}_{k_{\mathrm{U}}})\mathrm{U}. \tag{12b}$$

Matrices J^{U} and S^{U} provide all pairwise information between objects for 13FRI with respect to U. Let J^{V} and S^{V} be the corresponding matrices for V. 13FRI compares J^{U} and S^{U} with J^{V} and S^{V} to measure how much U and V agree with the membership assignment of each object pair. Let us elaborate these matrices.

$\mathrm{J}^{\mathrm{U}}_{i,j}$ and $\mathrm{S}^{\mathrm{U}}_{i,j}$ can be interpreted in several ways. For EHCs, $\mathrm{J}^{\mathrm{U}}_{i,j} = 1$ (implying $\mathrm{S}^{\mathrm{U}}_{i,j} = 0$) means that objects x_i and x_j belong to the same cluster in solution U, and $\mathrm{J}^{\mathrm{U}}_{i,j} = 0$ (implying $\mathrm{S}^{\mathrm{U}}_{i,j} = 1$) means that they belong to different clusters in U. In the EHC domain, J^{U} is the same matrix as that defined in Equation (6), and $\mathrm{S}^{\mathrm{U}}_{i,j} = 1 - \mathrm{J}^{\mathrm{U}}_{i,j}$.

Another interpretation can be provided for J^{U} and S^{U} in the FC domain. If one considers that an FC U produces probabilities of objects pertaining to clusters (e.g., as in EM solutions), i.e., $\mathrm{U}_{r,i}$ is the probability of object x_i belonging to the rth cluster, $\mathrm{J}^{\mathrm{U}}_{i,j}$ gives the probability of objects x_i and x_j belonging to the same cluster according to U, and $\mathrm{S}^{\mathrm{U}}_{i,j} = 1 - \mathrm{J}^{\mathrm{U}}_{i,j}$ gives the probability that they belong to different clusters according to U, assuming independence.

We also allow J^{U} and S^{U} to be defined for PCs in general (Section 5). Let us thus consider two other interpretations for J^{U} and S^{U} in the PC domain. Letting U be an NEHC, $\mathrm{J}^{\mathrm{U}}_{i,j}$ is the number of times x_i and x_j belong to the same cluster in U, and $\mathrm{S}^{\mathrm{U}}_{i,j}$ is the number of times x_i and x_j belong to different clusters in U. If U is a more general PC, we can say that $\mathrm{J}^{\mathrm{U}}_{i,j}$ is the possibility of x_i and x_j belonging to the same cluster in U, and $\mathrm{S}^{\mathrm{U}}_{i,j}$ is the possibility of x_i and x_j belonging to different clusters in U.

Despite the above multitude of interpretations, we understand that $\mathrm{J}^{\mathrm{U}}_{i,j}$ represents a degree of truthiness for the sentence "x_i and x_j belong to the same cluster", whereas $\mathrm{S}^{\mathrm{U}}_{i,j}$ yields a degree of falseness to the same sentence, according to the solution U. This reasoning

led us to redefine the pairing variables a, b, c, and d as follows:

$$\dot{a} \triangleq \sum_{i<j} \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}, \tag{13a}$$

$$\dot{b} \triangleq \sum_{i<j} \min\{\mathrm{J}^{\mathrm{U}}_{i,j} - \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}, \mathrm{S}^{\mathrm{V}}_{i,j} - \min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\}\}, \tag{13b}$$

$$\dot{c} \triangleq \sum_{i<j} \min\{\mathrm{J}^{\mathrm{V}}_{i,j} - \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}, \mathrm{S}^{\mathrm{U}}_{i,j} - \min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\}\}, \text{ and} \tag{13c}$$

$$\dot{d} \triangleq \sum_{i<j} \min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\}. \tag{13d}$$

Variables $\dot{a}$ and $\dot{d}$ measure the agreement between U and V with respect to the truthiness and falseness of sentence "x_i and x_j belong to the same cluster" for each pair of objects x_i and x_j; $\dot{b}$ and $\dot{c}$ measure the disagreement. For EHCs U and V, $\min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\} = 1$ means that x_i and x_j are clustered together in both clusterings. Conversely, $\min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\} = 1$ means that x_i and x_j belong to different clusters in both clusterings. In both cases, $\dot{a} + \dot{d}$ increases by 1. $\mathrm{J}^{\mathrm{U}}_{i,j} \neq \mathrm{J}^{\mathrm{V}}_{i,j}$ means that there is a disagreement between U and V regarding the pairing of x_i and x_j; it implies that $\min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\} = \min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\} = 0$ and increments $\dot{b} + \dot{c}$ by 1. This behavior recalls the descriptive definition of a, b, c, and d given in Section 3. Comparing the definitions in Equations (7) with those in Equations (13), $a = \dot{a}$, $b = \dot{b}$, $c = \dot{c}$, and $d = \dot{d}$ when comparing EHCs. Consequently, our similarity measure

$$\mathrm{13FRI}(\mathrm{U}, \mathrm{V}) \triangleq \frac{\dot{a} + \dot{d}}{\dot{a} + \dot{b} + \dot{c} + \dot{d}} \tag{14}$$

reduces to RI when U and V are EHCs.

Now, consider the more general context where U and V are FCs. We defined $\dot{a} + \dot{d}$ ($\dot{b} + \dot{c}$) to measure to what extent U and V agree (disagree) with each other regarding the object pairings. For example, the min operator in $\min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\}$ appears to provide a reasonable notion to what extent the solutions agree that x_i and x_j should not be clustered together. When the elements of J^{U} and J^{V} (or S^{U} and S^{V}) simultaneously show high or low values, there is a strong compatibility between U and V. This is reflected by how 13FRI was defined.

One may ask why $\dot{b}$ (and similarly for $\dot{c}$) was not defined as $\dot{b} \triangleq \sum_{i<j} \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\}$. The reason is that the amount $\min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}$ has already been used from $\mathrm{J}^{\mathrm{U}}_{i,j}$ and $\mathrm{J}^{\mathrm{V}}_{i,j}$ to establish the agreement between $\mathrm{J}^{\mathrm{U}}_{i,j}$ and $\mathrm{J}^{\mathrm{V}}_{i,j}$ in $\dot{a}$. Suppose that $\mathrm{J}^{\mathrm{U}}_{i,j} = \mathrm{S}^{\mathrm{U}}_{i,j} = \mathrm{J}^{\mathrm{V}}_{i,j} = \mathrm{S}^{\mathrm{V}}_{i,j} = x$. Let $\dot{a}_{i,j} \triangleq \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}$, and analogously define $\dot{b}_{i,j}$, $\dot{c}_{i,j}$, and $\dot{d}_{i,j}$. Without the subtractions in Equations (13b) and (13c), each variable $\dot{a}$, $\dot{b}$, $\dot{c}$, and $\dot{d}$ would be increased by x (i.e., $\dot{a}_{i,j} = \dot{b}_{i,j} = \dot{c}_{i,j} = \dot{d}_{i,j} = x$), meaning that U and V would have only 50% agreement regarding the placement of x_i and x_j, instead of 100%. This does not happen with the original formulation because all the information regarding the placement of x_i and x_j has been used in the definition of $\dot{a}_{i,j}$ and $\dot{d}_{i,j}$, and then nothing is left to the definition of $\dot{b}_{i,j}$ and $\dot{c}_{i,j}$. Figure 2 represents the values $\mathrm{J}^{\mathrm{U}}_{i,j} + \mathrm{S}^{\mathrm{U}}_{i,j} = 2x$ and $\mathrm{J}^{\mathrm{V}}_{i,j} + \mathrm{S}^{\mathrm{V}}_{i,j} = 2x$ by box heights. Parallel line orientations define the two types of filled areas regarding the information used

from the co-association matrices to determine $\dot{a}_{i,j}$ and $\dot{d}_{i,j}$. There is no space in the boxes (i.e., unused information) to fill regarding variables $\dot{b}_{i,j}$ and $\dot{c}_{i,j}$.

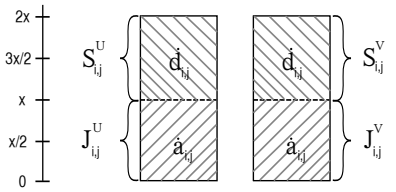


Figure 2: Graphical representation of a 13FRI evaluation where $\dot{b}_{i,j} = \dot{c}_{i,j} = 0$.

The 13FRI measure yields values in the continuous interval $[0, 1]$. It attains the maximum 1 whenever equivalent solutions are compared[9] and attains the minimum 0 only when U and V are EHCs and one of them has one cluster and the other has n clusters (Proposition 1 in Appendix). However, this last scenario is extreme and has little practical value (Vinh et al., 2009, 2010), making low 13FRI evaluations nearly impossible in practice. It is desirable that the entire interval $[0, 1]$ be useful, for better intuitiveness. This can be achieved by a similarity measure that takes values close to a constant α (α can always be turned into zero by a non-linear transformation: subtracting α from the evaluation and multiplying the result by a β that makes the maximum equals 1) when comparing random solutions (constant baseline). When a constant baseline exists and the user knows its value beforehand, one can compare the obtained evaluation to the baseline value and be more confident in his conclusions. The next section shows how 13FRI can be adjusted to assume values close to zero for randomly generated solutions.

4.1 Adjustment for Randomness

Suppose a measure assigns x to the similarity between two FCs U and V. How can we determine if x is not just a value from the random fluctuation inherent to the measure? A popular approach addresses this issue by subtracting the measure expectation from the measure and normalizing the result to 1 as a maximum (Hubert and Arabie, 1985; Albatineh et al., 2006; Vinh et al., 2009, 2010):

$$\mathrm{ASM}(\mathrm{U},\mathrm{V}) \triangleq \frac{\mathrm{SM}(\mathrm{U},\mathrm{V}) - \mathrm{E}[\mathrm{SM}]_{\mathrm{U},\mathrm{V}}}{\max\{\mathrm{SM}\} - \mathrm{E}[\mathrm{SM}]_{\mathrm{U},\mathrm{V}}}, \tag{15}$$

where SM is any similarity measure, $\mathrm{E}[\mathrm{SM}]_{\mathrm{U},\mathrm{V}}$ is its expectation given U and V, $\max\{\mathrm{SM}\}$ is the maximum of SM, and ASM is its adjusted version. ASM assumes values in the range $(-\infty, 1]$, and a positive value indicates that the similarity between U and V is greater than what one would expect from randomly chosen solutions. As Section 7.2 indicates for our corrected measures, this adjustment for chance can also make the measure unbiased in the number of clusters (Vinh et al., 2009, 2010).

To correct a measure for randomness, it is necessary to specify a null model according to which solutions are generated (Vinh et al., 2009, 2010). Given two FCs U and V, our

9. Note that $\mathrm{J}^{\mathrm{U}}_{i,j}$ and $\mathrm{S}^{\mathrm{U}}_{i,j}$ are independent of U row permutations. If U and V are equivalent clusterings, we have $\mathrm{J}^{\mathrm{U}}_{i,j} = \mathrm{J}^{\mathrm{V}}_{i,j}$ and $\mathrm{S}^{\mathrm{U}}_{i,j} = \mathrm{S}^{\mathrm{V}}_{i,j}$ $\forall i < j$. It implies that $\dot{b} = \dot{c} = 0$ and $\mathrm{13FRI}(\mathrm{U},\mathrm{V}) = 1$.

null model simultaneously produces two solutions from independent random permutations of the U and V columns. Let $\pi_1, \pi_2, \ldots, \pi_{n!}$ be every possible permutation of the numbers in $\mathbb{N}_{1,n}$, and define the function $\Gamma_{\pi_l}(\mathrm{U}) \triangleq [\mathrm{U}_{:,\pi_l(1)} \quad \mathrm{U}_{:,\pi_l(2)} \quad \ldots \quad \mathrm{U}_{:,\pi_l(n)}]$ that applies permutation π_l to matrix U.[10] A particular permutation π_l of U is chosen with probability $\mathrm{P}(\pi_l) \triangleq 1/n!$, and the permutations of U and V are considered independent events. We thus define $\mathrm{P}(\pi_l, \pi_q) \triangleq 1/(n!n!)$. The expectation of 13FRI according to our null model given U and V is

$$\mathrm{E}[13\mathrm{FRI}]_{\mathrm{U,V}} = \frac{1}{n!n!} \sum_{l,q=1}^{n!} 13\mathrm{FRI}(\Gamma_{\pi_l}(\mathrm{U}), \Gamma_{\pi_q}(\mathrm{V})). \tag{16}$$

Let $\dot{a}(\mathrm{J}^{\mathrm{U}}, \mathrm{J}^{\mathrm{V}}) \triangleq \sum_{i<j} \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}$ and $\dot{d}(\mathrm{S}^{\mathrm{U}}, \mathrm{S}^{\mathrm{V}}) \triangleq \sum_{i<j} \min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\}$. Because $\dot{a} + \dot{b} + \dot{c} + \dot{d}$ is a constant for the proposed null model (Corollary 1 in Appendix), we rewrite the expectation

$$\mathrm{E}[13\mathrm{FRI}]_{\mathrm{U,V}} = (\dot{a} + \dot{b} + \dot{c} + \dot{d})^{-1}(\mathrm{E}[\dot{a}]_{\mathrm{U,V}} + \mathrm{E}[\dot{d}]_{\mathrm{U,V}}), \tag{17}$$

where

$$\begin{aligned}
\mathrm{E}[\dot{a}]_{\mathrm{U,V}} &= \frac{1}{n!n!} \sum_{l,q=1}^{n!} \dot{a}(\mathrm{J}^{\Gamma_{\pi_l}(\mathrm{U})}, \mathrm{J}^{\Gamma_{\pi_q}(\mathrm{V})}) \\
&= \frac{1}{n!n!} \sum_{l,q=1}^{n!} \sum_{i_1<j_1} \min\{\mathrm{J}^{\mathrm{U}}_{\pi_l(i_1),\pi_l(j_1)}, \mathrm{J}^{\mathrm{V}}_{\pi_q(i_1),\pi_q(j_1)}\} \\
&= \frac{2(n-2)!}{n!n!} \sum_{q=1}^{n!} \sum_{i_1<j_1} \sum_{i_2<j_2} \min\{\mathrm{J}^{\mathrm{U}}_{i_2,j_2}, \mathrm{J}^{\mathrm{V}}_{\pi_q(i_1),\pi_q(j_1)}\} \\
&= \frac{2(n-2)!2(n-2)!}{n!n!} \sum_{i_1<j_1} \sum_{i_2<j_2} \sum_{i_3<j_3} \min\{\mathrm{J}^{\mathrm{U}}_{i_2,j_2}, \mathrm{J}^{\mathrm{V}}_{i_3,j_3}\} \\
&= \frac{4}{n^2(n-1)^2} \sum_{i_1<j_1} \sum_{i_2<j_2} \sum_{i_3<j_3} \min\{\mathrm{J}^{\mathrm{U}}_{i_2,j_2}, \mathrm{J}^{\mathrm{V}}_{i_3,j_3}\} \\
&= \frac{2}{n(n-1)} \sum_{i_2<j_2} \sum_{i_3<j_3} \min\{\mathrm{J}^{\mathrm{U}}_{i_2,j_2}, \mathrm{J}^{\mathrm{V}}_{i_3,j_3}\}
\end{aligned} \tag{18}$$

and, analogously,

$$\mathrm{E}[\dot{d}]_{\mathrm{U,V}} = \frac{2}{n(n-1)} \sum_{i_2<j_2} \sum_{i_3<j_3} \min\{\mathrm{S}^{\mathrm{U}}_{i_2,j_2}, \mathrm{S}^{\mathrm{V}}_{i_3,j_3}\}. \tag{19}$$

Following the framework of Equation (15), the adjusted frand index is

$$13\mathrm{AFRI}(\mathrm{U,V}) \triangleq \frac{13\mathrm{FRI}(\mathrm{U,V}) - \mathrm{E}[13\mathrm{FRI}]_{\mathrm{U,V}}}{1 - \mathrm{E}[13\mathrm{FRI}]_{\mathrm{U,V}}}. \tag{20}$$

10. $\mathrm{U}_{:,i}$ is the ith column of U.

13AFRI attains its maximum 1 in the same way as 13FRI (i.e., whenever two equivalent clusterings are compared) and is 0 when the measure equals its expected value, under the null model. 13AFRI can display negative evaluations, which mean that the compared clusterings are more dissimilar than expected if they were independently generated. Its minimum is not fixed anymore and is given by $-\mathrm{E[SM]_{U,V}}/(\max\{\mathrm{SM}\} - \mathrm{E[SM]_{U,V}})$.

Given two EHCs U and V, we have $\mathrm{13AFRI(U,V)} = \mathrm{ARI(U,V)}$ (Proposition 3 in Appendix). In other words, 13AFRI reduces to ARI in the EHC domain. This indicates the appropriateness of the null model for 13AFRI, which can also be further extended to PCs (as Section 5 shows).

4.2 Discussion

13FRI could also be applied to PCs. In this case, however, 13FRI would not provide reasonable evaluations in some scenarios where per-object membership totals (i.e., column-wise sums of the clustering matrix) varies among solutions. Let U be an FC and recall that an FC is also a PC. The result of multiplying U by a scalar $x \in (0,1)$ is also a PC matrix, where the per-object membership total of each object is decreased. Notice that we have $\mathrm{13FRI(U,U)} = \mathrm{13FRI(U},x\mathrm{U)} = \mathrm{13AFRI(U,U)} = \mathrm{13AFRI(U},x\mathrm{U)} = 1$ for any $x \in (0,1]$. This happens because $\mathrm{J}^{x\mathrm{U}}_{i,j} = \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{x\mathrm{U}}_{i,j}\}$ and $\mathrm{S}^{x\mathrm{U}}_{i,j} = \min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{x\mathrm{U}}_{i,j}\}$, making variables $\dot{b}$ and $\dot{c}$ (Equations 13b and 13c) equal to zero.

Let us analyze another problematic scenario by considering the following matrices:

$$\mathrm{U} \triangleq \begin{pmatrix} 0.8 & 0.4 \\ 0.4 & 0.8 \end{pmatrix} \quad \text{and} \quad \mathrm{V} \triangleq \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}.$$

Note that U is a PC more general than an FC. We have $\mathrm{J}^{\mathrm{U}}_{1,2} = 0.64$, $\mathrm{S}^{\mathrm{U}}_{1,2} = 0.8$, $\mathrm{J}^{\mathrm{V}}_{1,2} = 0.48$, and $\mathrm{S}^{\mathrm{V}}_{1,2} = 0.52$. The heights of the first and second boxes in Figure 3 correspond to the values $\mathrm{J}^{\mathrm{U}}_{1,2}+\mathrm{S}^{\mathrm{U}}_{1,2} = 1.44$ and $\mathrm{J}^{\mathrm{V}}_{1,2}+\mathrm{S}^{\mathrm{V}}_{1,2} = 1$, respectively. The boxes are divided by horizontal dashed lines, creating two parts that correspond to the $\mathrm{J}^{\mathrm{U}}_{1,2}$ and $\mathrm{S}^{\mathrm{U}}_{1,2}$ ($\mathrm{J}^{\mathrm{V}}_{1,2}$ and $\mathrm{S}^{\mathrm{V}}_{1,2}$) values. The values of $\dot{a} = 0.48$ and $\dot{d} = 0.52$ are illustrated by the filled areas, and the remaining variables $\dot{b}$ and $\dot{c}$ equal zero. There is an empty space of height $\mathrm{J}^{\mathrm{U}}_{1,2}+\mathrm{S}^{\mathrm{U}}_{1,2}-(\mathrm{J}^{\mathrm{V}}_{1,2}+\mathrm{S}^{\mathrm{V}}_{1,2}) = 0.44$ in the first box, which 13FRI ignores. We could increase $\mathrm{J}^{\mathrm{U}}_{1,2}$ and $\mathrm{S}^{\mathrm{U}}_{1,2}$ by any amount that 13FRI would still yield the same score. A reasonable measure for PCs should decrease the score proportionally to the unmatched amount. The next section proposes modifying 13FRI to address this issue.

5. Grand Index

Let $\mathrm{T^U} \triangleq \mathrm{J^U} + \mathrm{S^U}$ and $\mathrm{M} \triangleq \max\{\mathrm{T^U}, \mathrm{T^V}\}$.[11] A new variable

$$\dot{e} \triangleq \max\left\{ \sum_{i<j} \left(\mathrm{M}_{i,j} - \mathrm{T}^{\mathrm{U}}_{i,j}\right), \sum_{i<j} \left(\mathrm{M}_{i,j} - \mathrm{T}^{\mathrm{V}}_{i,j}\right) \right\} \tag{21}$$

11. $\mathrm{M} = \max\{\mathrm{T^U}, \mathrm{T^V}\}$ means that $\mathrm{M}_{i,j} = \max\{\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j}\}$ for all i, j.

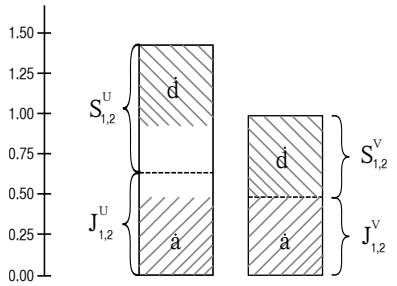


Figure 3: Graphical representation of the problem using 13FRI when the compared clustering matrices have different column-wise sums.

is introduced in 13FRI to give rise to the grand index:

$$\text{13GRI}(\text{U},\text{V}) \triangleq \frac{\dot{a}+\dot{d}}{\dot{a}+\dot{b}+\dot{c}+\dot{d}+\dot{e}}. \tag{22}$$

Given two objects x_i and x_j, $\text{M}_{i,j} - \text{T}^{\text{U}}_{i,j}$ describes how much $\text{T}^{\text{V}}_{i,j}$ exceeds $\text{T}^{\text{U}}_{i,j}$. In Figure 3, $\text{M}_{i,j} - \text{T}^{\text{V}}_{i,j} = 0.44$, which equals the height of the empty space in the first box. Proposition 5 in Appendix allows us to rewrite Equation (22) as

$$\text{13GRI}(\text{U},\text{V}) = \frac{\sum_{i<j} \min\{\text{J}^{\text{U}}_{i,j}, \text{J}^{\text{V}}_{i,j}\} + \sum_{i<j} \min\{\text{S}^{\text{U}}_{i,j}, \text{S}^{\text{V}}_{i,j}\}}{\max\{\sum_{i<j} \text{T}^{\text{U}}_{i,j}, \sum_{i<j} \text{T}^{\text{V}}_{i,j}\}}.$$

If U and V are FCs, $\text{T}^{\text{U}}_{i,j} = \text{T}^{\text{V}}_{i,j} = 1$, $\dot{a}+\dot{b}+\dot{c}+\dot{d}+\dot{e} = \max\{\sum_{i<j} \text{T}^{\text{U}}_{i,j}, \sum_{i<j} \text{T}^{\text{V}}_{i,j}\} = n(n-1)/2$, and 13GRI reduces to 13FRI. As in 13FRI, 13GRI attains its maximum 1 whenever the compared PCs U and V are equivalent solutions.[12]

Adopting the same null model proposed in Section 4.1, and realizing that $\dot{a}+\dot{b}+\dot{c}+\dot{d}+\dot{e}$ is constant for this model (Corollary 2 in Appendix), we have $\text{E}[\text{13GRI}]_{\text{U,V}} = (\dot{a}+\dot{b}+\dot{c}+\dot{d}+\dot{e})^{-1}(\text{E}[\dot{a}]_{\text{U,V}} + \text{E}[\dot{d}]_{\text{U,V}})$. The adjusted 13GRI is then given by

$$\text{13AGRI}(\text{U},\text{V}) \triangleq \frac{\text{13GRI}(\text{U},\text{V}) - \text{E}[\text{13GRI}]_{\text{U,V}}}{1 - \text{E}[\text{13GRI}]_{\text{U,V}}}. \tag{23}$$

Similarly to 13AFRI, 13AGRI attains its maximum 1 in the same way as 13GRI and is 0 when the measure equals its expected value. Section 7.2 shows that 13AGRI can indeed exhibit a constant baseline close to zero for randomly generated EHC, FC, NEHC, and PC solutions, even when the null model is clearly violated.

6. Computational Complexity and Implementation

Let I_{k_U} be the k_U-by-k_U identity matrix and $\mathbb{1}_{k_\text{U}}$ the k_U-by-k_U matrix with 1 in each entry. There are $O(n^2 k_\text{U})$ computational steps to calculate $\text{J}^{\text{U}} = \text{U}^{\text{T}}\text{U}$ and $\text{S}^{\text{U}} = \text{U}^{\text{T}}(\mathbb{1}_{k_\text{U}} - \text{I}_{k_\text{U}})\text{U}$,

12. As in the 13FRI case, we have $\text{J}^{\text{U}}_{i,j} = \text{J}^{\text{V}}_{i,j}$ and $\text{S}^{\text{U}}_{i,j} = \text{S}^{\text{V}}_{i,j}$ $\forall i < j$ whenever U and V are equivalent clusterings. Thus, $\text{T}^{\text{U}}_{i,j} = \text{T}^{\text{V}}_{i,j}$ $\forall i < j$, making $\dot{a}$ and $\dot{d}$ the only possible non-null terms.

and $O(n^2)$ steps to calculate $\mathrm{M} = \max\{\mathrm{T}^{\mathrm{U}}, \mathrm{T}^{\mathrm{V}}\}$. Variables $\dot{a}$, $\dot{b}$, $\dot{c}$, $\dot{d}$, and $\dot{e}$ require $O(n^2)$ steps because of the pairwise summations $\sum_{i<j}$ in their formulas (Equations 13 and 21). 13FRI and 13GRI thus require $O(n^2(k_{\mathrm{U}} + k_{\mathrm{V}}))$ operations. Calculation of EHC pair-based measures generally requires $O(nk_{\mathrm{U}}k_{\mathrm{V}})$ steps due to the contingency matrix $\mathrm{N} = \mathrm{U}\mathrm{V}^{\mathrm{T}}$ computation. The possibly higher 13FRI and 13GRI complexity is the price one may have to pay for a more general measure.

Equations (18) and (19) might suggest that 13AFRI (and 13AGRI) requires $O(n^4)$ computational steps, making its computation infeasible for most practical scenarios. Fortunately, the min operator allows us to reduce the computational complexity of Equations (18) and (19) to $O(n^2 \log n)$ steps. To examine how that can be accomplished, suppose that $\mathrm{J}^{\mathrm{U}}_{1,2} \leq \mathrm{J}^{\mathrm{V}}_{i,j}$ for all $i < j$ $(i, j \in \mathbb{N}_{1,n})$ as a special case and as a didactic example. We have $\sum_{i<j} \min\{\mathrm{J}^{\mathrm{U}}_{1,2}, \mathrm{J}^{\mathrm{V}}_{i,j}\} = \mathrm{J}^{\mathrm{U}}_{1,2} n(n-1)/2$ computable in constant time, reducing the total computational cost. Let us consider the general case for calculating $\mathrm{E}[\dot{a}]_{\mathrm{U,V}}$ (Equation 18). Define

$$\mathbf{1}^{i_2,j_2}_{i_1,j_1} \triangleq \begin{cases} 1 & \text{if } \mathrm{J}^{\mathrm{U}}_{i_1,j_1} \leq \mathrm{J}^{\mathrm{V}}_{i_2,j_2} \\ 0 & \text{otherwise} \end{cases}.$$

Equation (18) can be rewritten as

$$\begin{aligned} \frac{n(n-1)}{2}\mathrm{E}[\dot{a}]_{\mathrm{U,V}} &= \sum_{i_1<j_1}\sum_{i_2<j_2} \min\{\mathrm{J}^{\mathrm{U}}_{i_1,j_1}, \mathrm{J}^{\mathrm{V}}_{i_2,j_2}\}\mathbf{1}^{i_2,j_2}_{i_1,j_1} + \sum_{i_2<j_2}\sum_{i_1<j_1} \min\{\mathrm{J}^{\mathrm{U}}_{i_1,j_1}, \mathrm{J}^{\mathrm{V}}_{i_2,j_2}\}(1 - \mathbf{1}^{i_2,j_2}_{i_1,j_1}) \\ &= \sum_{i_1<j_1} \mathrm{J}^{\mathrm{U}}_{i_1,j_1} \sum_{i_2<j_2} \mathbf{1}^{i_2,j_2}_{i_1,j_1} + \sum_{i_2<j_2} \mathrm{J}^{\mathrm{V}}_{i_2,j_2} \sum_{i_1<j_1} (1 - \mathbf{1}^{i_2,j_2}_{i_1,j_1}). \end{aligned} \tag{24}$$

The calculation of $\mathrm{E}[\dot{d}]_{\mathrm{U,V}}$ (Equation 19) is analogous; the only difference lies in using S^{U} and S^{V} instead of J^{U} and J^{V}.

The above strategy can be applied efficiently by first rearranging the upper triangular parts of J^{U} and J^{V} into vectors x and y, respectively, and sorting the resulting vectors.[13] Algorithm 1 shows an implementation of the above strategy, where the first and second terms of the right-hand side of Equation (24) are calculated by the loops in Steps 7 and 15, respectively.

The most demanding step of Algorithm 1 in terms of computational time is Step 4, which sorts two vectors of size $n(n-1)/2$ in $O(n^2 \log n)$ steps using, for example, the heap sort algorithm. 13AGRI and 13AFRI thus require $O(n^2(k_{\mathrm{U}} + k_{\mathrm{V}} + \log n))$ computational steps.

7. Experiments

It is a common practice to compare the accuracy of clustering algorithms by measuring how similar their resulting clusterings are to a reference solution. The algorithm that generated clusterings more similar to the reference solution is then regarded as the most accurate.

13. The upper triangular part of $\mathrm{J}^{\mathrm{U}}_{i,j}$ can be rearranged as follows: $\mathrm{x}_{\pi(i,j)} \triangleq \mathrm{J}^{\mathrm{U}}_{i,j}$ $(\forall i < j)$, where $\pi(i,j) \triangleq j - i + \sum_{t=1}^{i-1}(n-t) = j - i(1+i)/2 + n(i-1)$.

Algorithm 1 Compute $\mathrm{E}[\dot{a}]_{\mathrm{U,V}}$

Represent the upper triangular part of J^{U} into vector x
Represent the upper triangular part of J^{V} into vector y
$m \leftarrow n(n-1)/2$ {size of vectors x and y}
Sort x and y in increasing order
$\mathrm{E}[\dot{a}]_{\mathrm{U,V}} \leftarrow 0$
$i, j \leftarrow m, m$
while $i > 0$ **do**
  while $j > 0$ **and** $\mathrm{x}_i \leq \mathrm{y}_j$ **do**
    $j \leftarrow j - 1$
  end while
  $\mathrm{E}[\dot{a}]_{\mathrm{U,V}} \leftarrow \mathrm{E}[\dot{a}]_{\mathrm{U,V}} + (m - j) * \mathrm{x}_i$
  $i \leftarrow i - 1$
end while
$i, j \leftarrow m, m$
while $j > 0$ **do**
  while $i > 0$ **and** $\mathrm{x}_i > \mathrm{y}_j$ **do**
    $i \leftarrow i - 1$
  end while
  $\mathrm{E}[\dot{a}]_{\mathrm{U,V}} \leftarrow \mathrm{E}[\dot{a}]_{\mathrm{U,V}} + (m - i) * \mathrm{y}_j$
  $j \leftarrow j - 1$
end while
$\mathrm{E}[\dot{a}]_{\mathrm{U,V}} \leftarrow \mathrm{E}[\dot{a}]_{\mathrm{U,V}}/m$

A measure must somehow adequately evaluate the similarity between the compared solutions. Section 7.1 follows this idea and compares 34 measures by applying them to evaluate solutions with different numbers of clusters produced by different clustering algorithms. This comparison is done by considering the first three properties proposed in Section 2: maximum, discriminant, and contrast. Synthetic data sets were generated according to the cluster types that these algorithms search for (e.g., it is well-known that k-means (MacQueen, 1967) tends to produce spherical-like clusters), and the reference solution for each data set was defined by applying the corresponding clustering algorithm with a well-tuned initial solution. In this scenario is then expected that the dissimilarity between the generated and reference solutions will reflect the difference in the numbers of clusters.

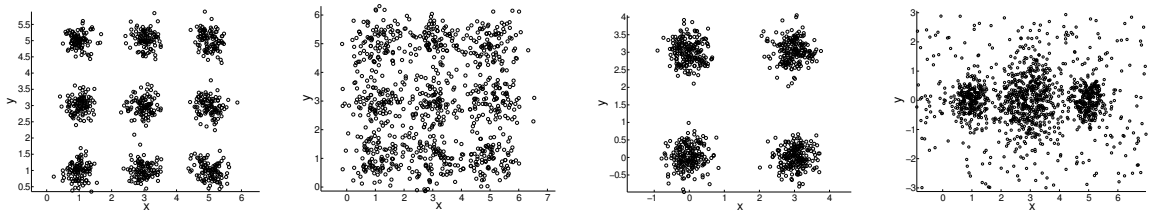
In a different scenario, Section 7.2 compares the measures when evaluating randomly generated solutions, by assessing the measures according to the baseline property proposed in Section 2. A measure should display a uniform evaluation across the range of numbers of clusters because any resemblance between the compared solutions is only due to chance.

Section 7.3 assesses the 13AGRI evaluation validity for FCs in 14 real data sets, and Section 7.4 uses 13AGRI as a stability statistic for estimating the number of clusters in five real data sets.

Because 13GRI (13AGRI) is more general and becomes equivalent to 13FRI (13AFRI) when applied to FCs, we only show the results of 13GRI (13AGRI).

7.1 Measuring the Similarity Between Clusterings

We evaluated the measures in four synthetic data sets (Figures 4), each suitable for one of the following clustering types: EHC, FC, NEHC, and PC. The DEHC data set (Figure 4(a)) has nine well-separated clusters, whereas the DFC data set (Figure 4(b)) has nine overlapping clusters. In both data sets, the clusters were generated using Gaussian distributions with equal variances and no correlation between the attributes. The DNEHC data set (Figure 4(c)) has four clusters, but they reduce to two clusters when projected to a single axis.[14] We generated the DPC data set (Figure 4(d)) to resemble a synthetic one (Zhang and Leung, 2004) with noise added.



(a) Data set for EHC (DEHC).

(b) Data set for FC (DFC).

(c) Data set for NEHC (DNEHC).

(d) Data set for PC (DPC).

Figure 4: Data set for each clustering type.

Different clustering algorithms were employed for each data set, appropriate for the corresponding clustering type as follows: k-means for DEHC, fuzzy c-means (FCM) and expectation maximization for Gaussian mixtures (EMGM) (Dempster et al., 1977) for DFC, SUBCLU (Kailing et al., 2004) for DNEHC, and improved possibilistic c-means 2 (IPCM2) (Zhang and Leung, 2004) for DPC. The FCM and IPCM2 exponent m was set to 2 (which is commonly adopted in the literature), the SUBCLU parameter *minpts* was set to 5, and the Euclidean norm was adopted; this same configuration was used in all the experiments reported in this work. The reference solution for the combination of data set and clustering algorithm (i.e., (DEHC, k-means), (DFC, FCM), (DFC, EMGM), (DNEHC, SUBCLU), and (DPC, IPCM2)) was produced by applying the clustering algorithm with the right number of clusters (or a well-tuned epsilon for SUBCLU), and the result was analyzed to ensure that the solution could be considered ideal in the clustering space sought by the corresponding algorithm. For example, we applied k-means to DEHC with $k = 9$ clusters, using the means of the Gaussian distributions (used to generate the clusters) as the initial centroids. The final solution had virtually the same initial centroids, corroborating the validity of the obtained solution.

It is worth noting that we are not suggesting that the considered clustering algorithms are not suitable for the data sets to which they have not been applied to. For example, FCM can easily find the clustering structure in DEHC, as well as IPCM2 can find the clustering structure in DFC. What is most important is that the data set has a clustering structure suitable for the clustering algorithm being applied.

14. The other data sets could have a similar interpretation as well. However, we only consider subspaces in this specific data set.

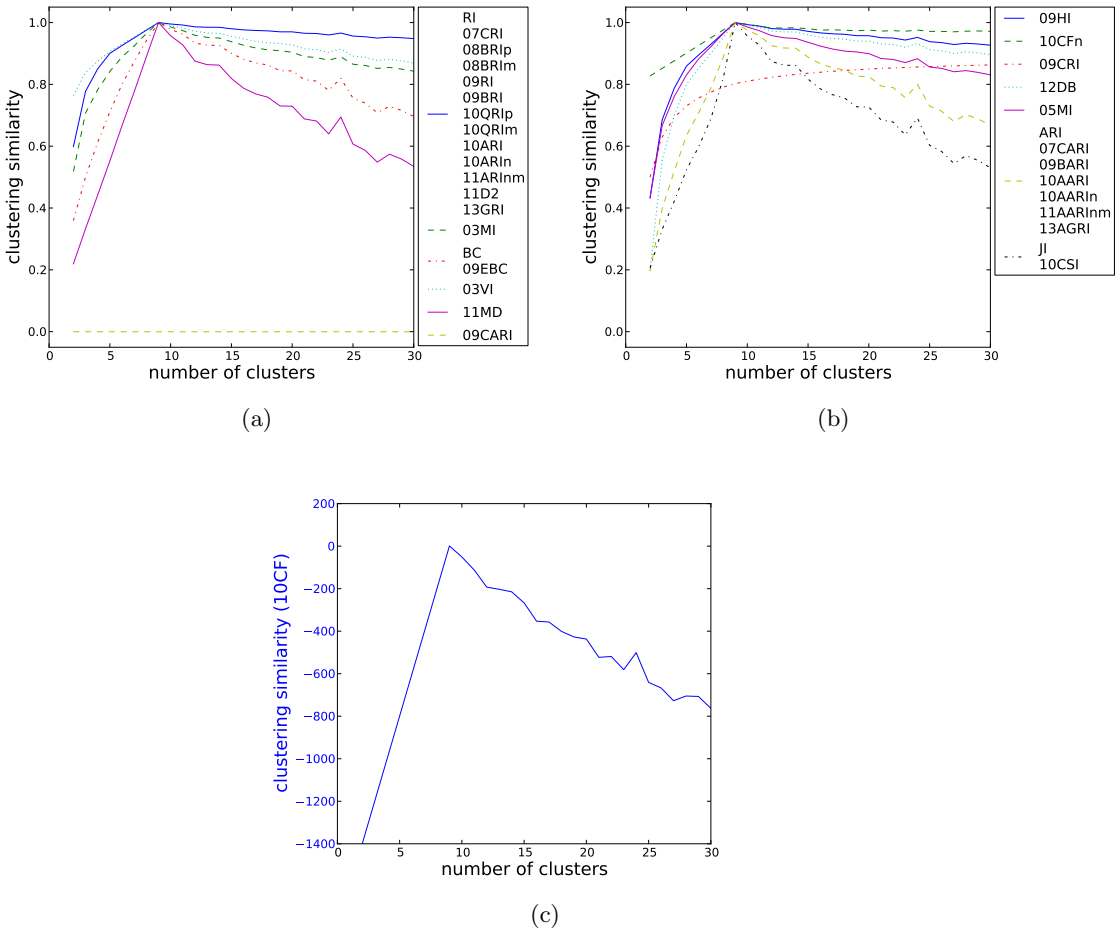


Figure 5: EHC measure evaluations of k-means solutions for the DEHC data set.

The algorithms k-means, FCM, EMGM, and IPCM2 were applied 30 times for each number of clusters $k \in \{2, 3, \ldots, \sqrt{n}\}$ (the literature commonly adopts the upper threshold $\sqrt{n}$ as a rule of thumb (Pal and Bezdek, 1995; Pakhira et al., 2005)), and SUBCLU was applied 30 times for each epsilon in the range $\{0.1, 0.2, \ldots, 5.0\}$. The measures were applied to each solution, and only the highest (which means "the best") values attained in each k or epsilon for a given measure were retained to generate the plots in Figures 5, 6, 7, 8, and 9. We opted to plot the highest values instead of averages because we are interested in the solutions that are as close as possible to the reference one, for a given number of clusters (or epsilon), and to make the results as independent as possible to the stochastic nature of the algorithms. Measures showing the same values were joined and represented by a single curve, and multiple figures for the same experiments were plotted for visualization purposes.

Figure 5 shows that most generalized measures displayed the same results as RI or ARI, when evaluating EHCs. This is expected because most of these measures were defined

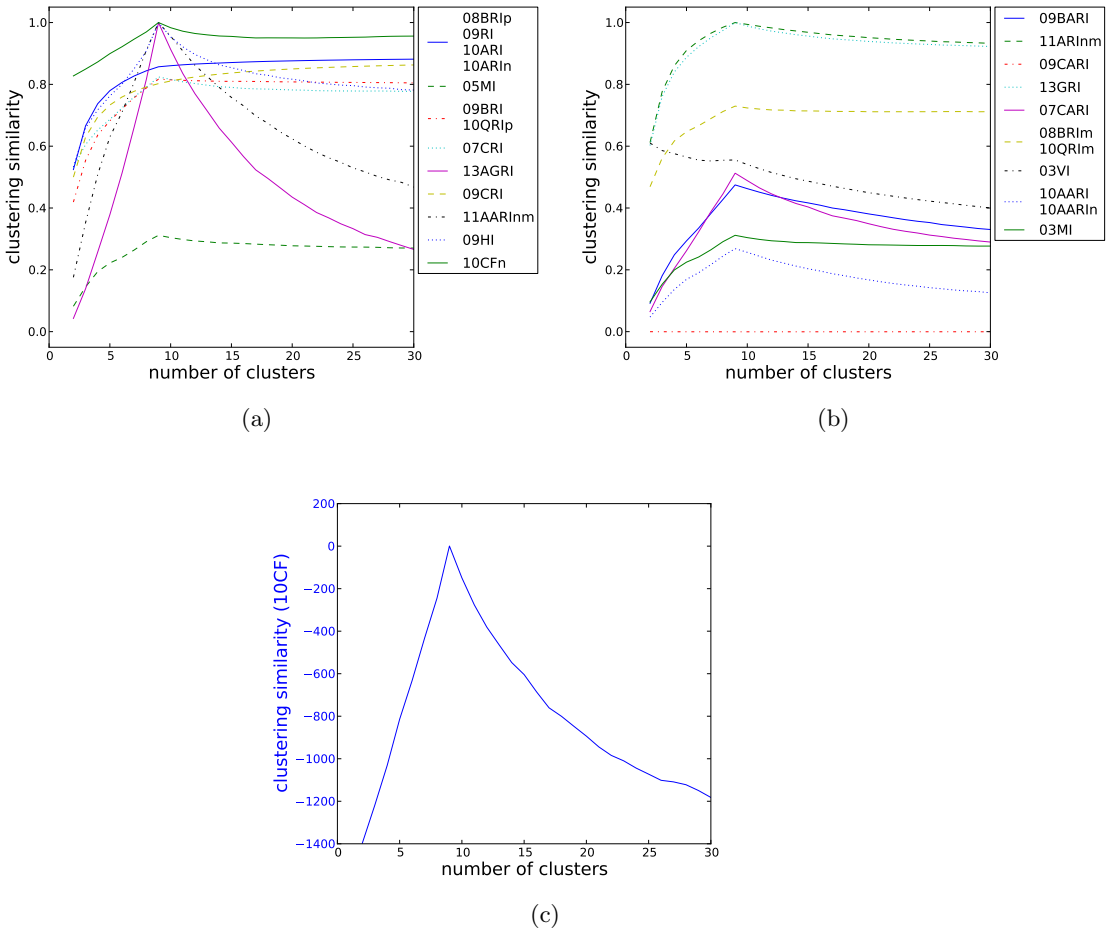


Figure 6: FC measure evaluations of FCM solutions for the DFC data set.

by extending the variables behind the RI or ARI formulations. For example, the 07CRI measure is a fuzzy version of RI in which the pairing variables a, b, c, and d were defined using fuzzy sets. When applied to EHCs, 07CRI reduces to RI (Campello, 2007). RI, 09HI, 10CFn, 12DB, and the measures that showed the same results as RI were weakly affected by a positive difference between the obtained and the true numbers of clusters. RI is equal to 1 and 0.94 for the solutions with 9 and 30 clusters, respectively, which represents less than 10% of its total range $[0, 1]$. This weak responsiveness to the number of clusters makes it difficult to decide whether the solution at hand is really good or not (weak contrast property). 09CRI exhibited an increasing evaluation across the numbers of clusters, and 09CARI produced scores close to zero only. In fact, 09CARI resulted in evaluations close to zero for each scenario in this section. Conversely, JI, ARI, BC, 09EBC, 10CSI, 11MD, and the measures that showed the same results as ARI (including 13AGRI proposed here) exhibited a steady decrease for high numbers of clusters. We believe that this more prominent responsiveness

to differences in the clusterings is more intuitively appealing. 10CF (Figure 5(c)) attained the maximum 1 for the right number of clusters.

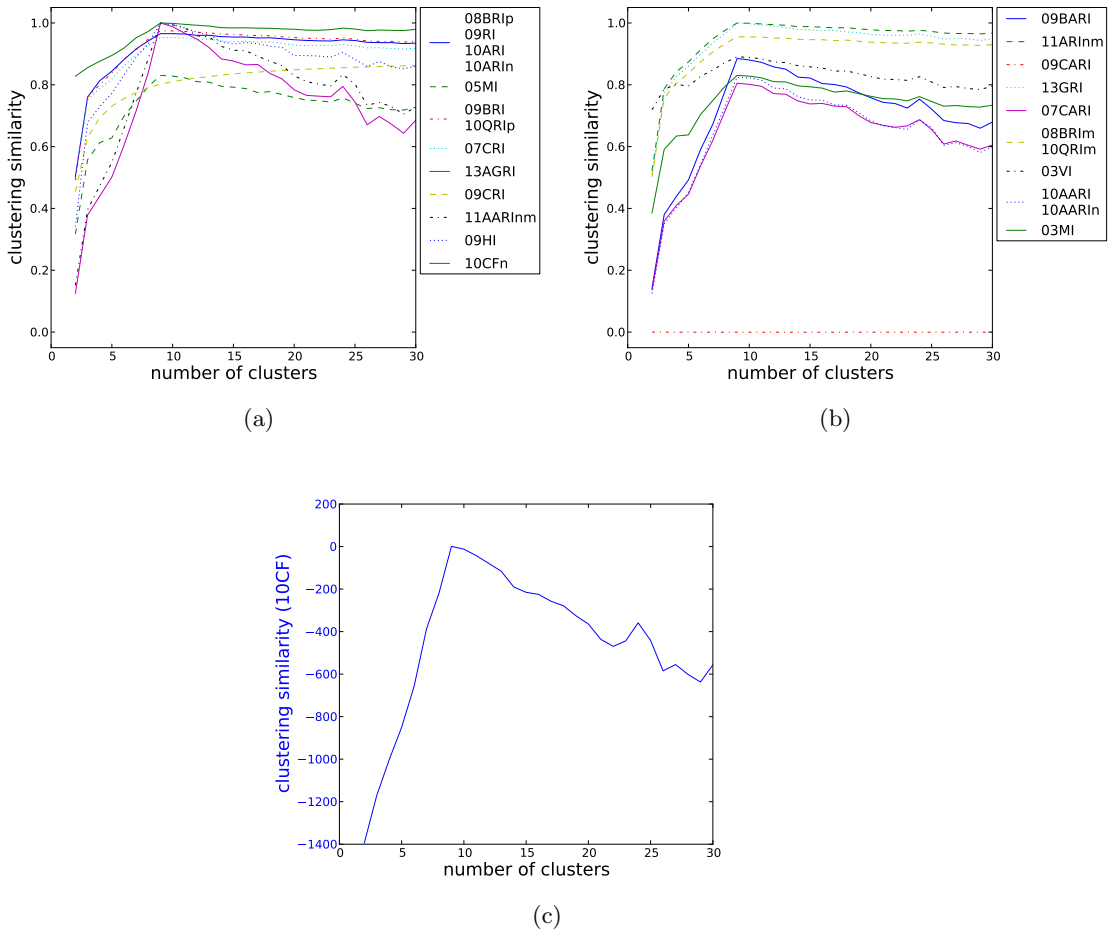


Figure 7: FC measure evaluations of EMGM solutions for the DFC data set.

Figure 6 shows FC measure evaluations of FCM solutions for the DFC data set. Only 13AGRI and 11AARInm provided both the maximum value 1 for the true number of clusters and showed steady decreasing evaluations over the positive increase in the difference between the obtained and true numbers of clusters. 09HI was 1 for the true number of clusters, but it showed an asymptotic-like curve for high numbers of clusters. 03VI, 08BRIp, 09RI, 09CRI, 09CARI, 10ARI, and 10ARIn could not indicate the reference solution.

Figure 7 displays EMGM solution evaluations for the DFC data set. 07CRI, 08BRIp, 08BRIm, 09CRI, 09CARI, 09RI, 09BRI, 10QRIp, 10QRIm, 10ARI, 10ARIn, and 11ARInm could not indicate the true number of clusters. 09HI, 10CFn, 11AARInm, 13GRI, and 13AGRI attained their maxima 1 for the right number of clusters. However, 10CFn and 13GRI showed little to no evaluation change over the solutions with number of clusters greater than $k^* = 9$ (low contrast). 10CF attained 0.92 for the right number of clusters.

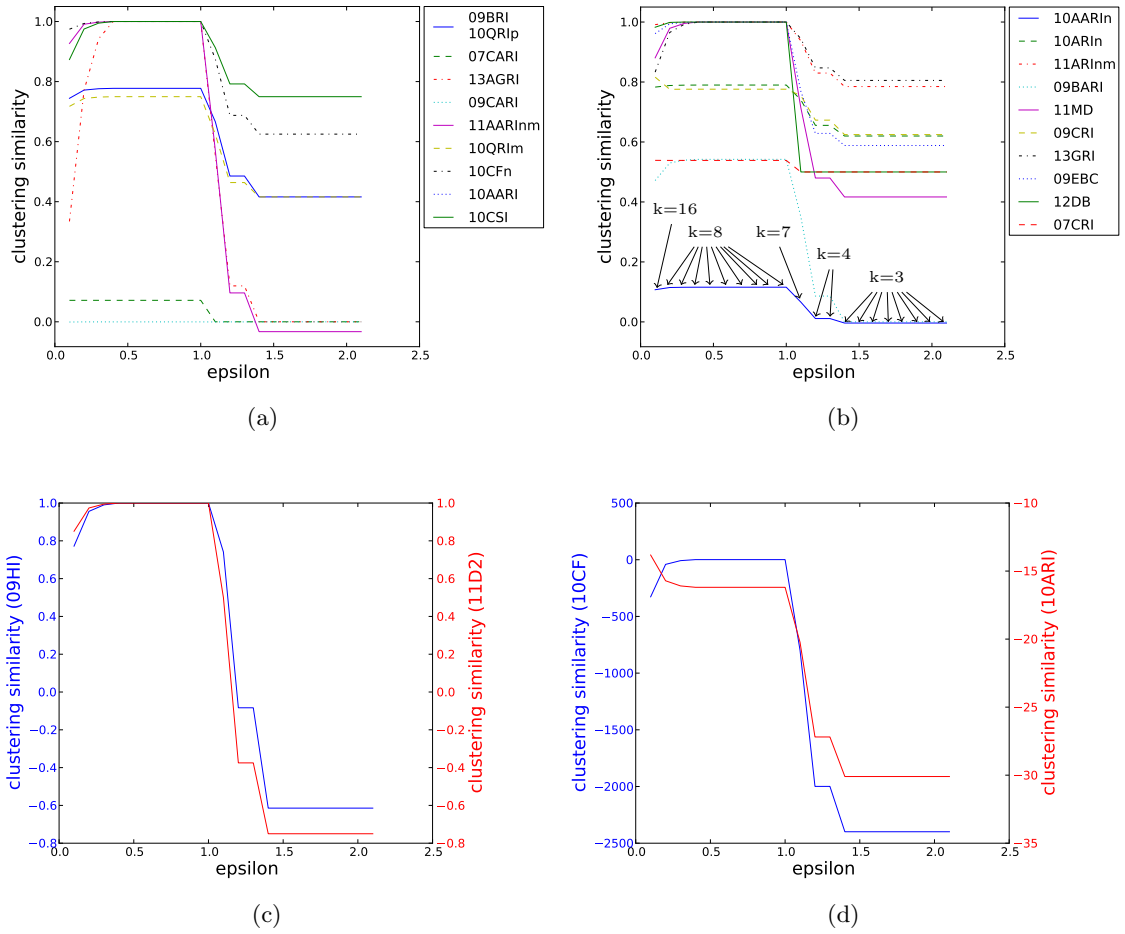


Figure 8: NEHC measure evaluations of SUBCLU solutions for the DNEHC data set.

Figure 8, in which NEHCs are evaluated, shows only the range $\{0.1, 0.2, \ldots, 2.1\}$ of epsilons, as the results from 1.4 to 5.0 are identical. The reference solution has 8 clusters: 4 from data on the plane, 2 from data projected onto the x axis, and 2 from data projected onto the y axis (Figure 4(c)). Figure 8(b) indicates the number of clusters found for each epsilon. SUBCLU generates the reference solution only for the epsilons from 0.4 to 1.0 (we know this by inspection), and most measures yield the highest score in this interval. 07CRI, 09CRI, 10ARI, and 10AARI judged the solution with an epsilon equal to 0.1 to be the best one. Most of the measures identified the correct solutions, but only 09EBC, 09HI, 10CSI, 10CF, 10CFn, 11AARInm, 11MD, 11D2, 13GRI, and 13AGRI attained their maxima 1 for these solutions. 11AARInm and 13AGRI rapidly approached zero for non-optimal epsilons.

In Figure 9, 13GRI and 13AGRI exhibited a steep fall in the evaluations and a peak 1 at the true number of clusters. The DPC data set has only 3 clusters, while the others have 9 (DEHC and DFC) or 8 (DNEHC) clusters. A steeper curve is therefore expected. 07CRI, 09HI, 09BRI, 10QRIp, 10QRIm, 10ARI, 10ARIn, and 11ARInm provided high evaluations

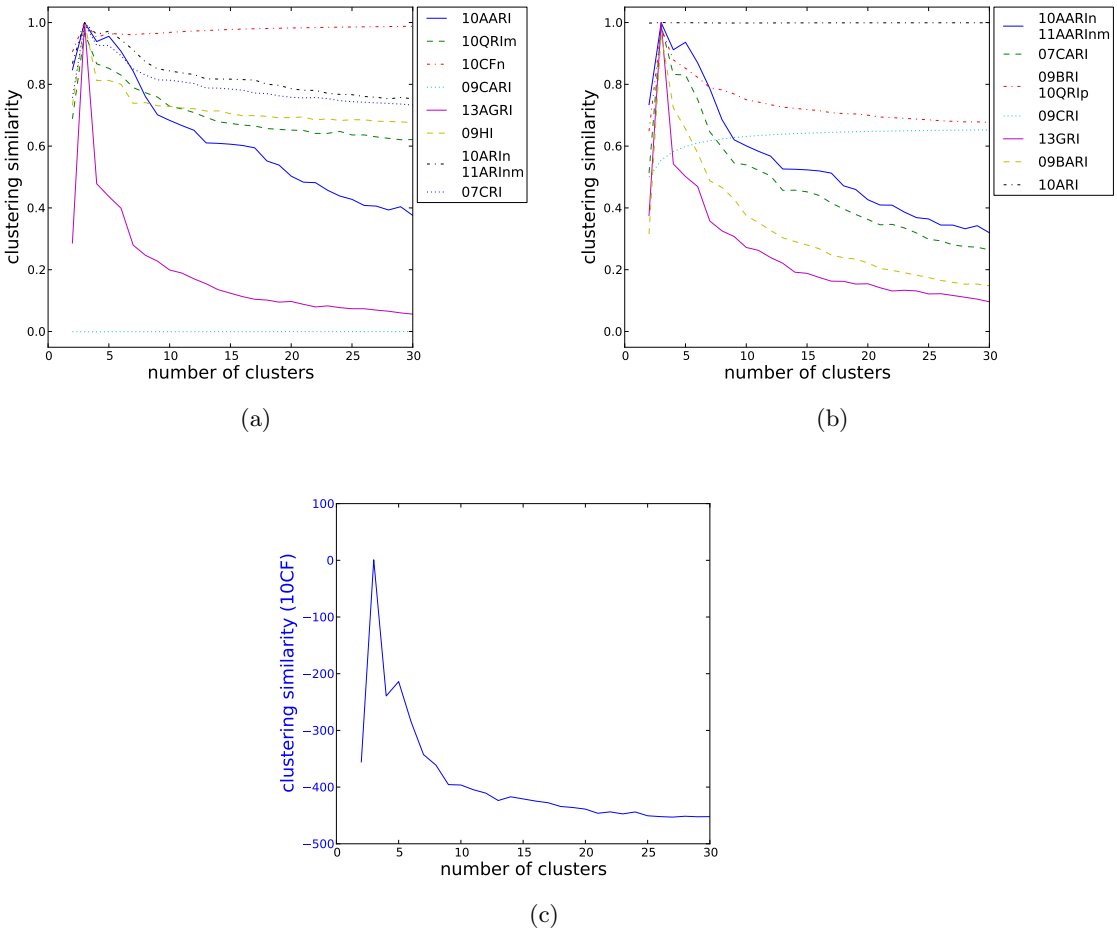


Figure 9: PC measure evaluations of IPCM2 solutions for the DPC data set.

for a wide range of numbers of clusters. Measures 10ARI and 09CARI could not discriminate between the solutions, and 09CRI could not indicate the true number of clusters. 10CFn showed an increasing evaluation for solutions with number of clusters greater than $k = 5$. 10CF indicated the right number of clusters in Fig 9(c), though not evaluating it as the maximum 1 (it was evaluated as 0.92).

Table 3 summarizes the results by indicating with "k*" the measures that identified the reference clustering (discriminant property) and "1" the measures that attained their maxima for the reference solution (maximum property). 09HI, 10CFn, 11AARInm, 13GRI, and 13AGRI are the only measures that displayed the above properties for each scenario. However, 09HI, 10CFn, and 13GRI presented a poor sensitivity to solution variations in most of the cases (e.g., Figures 5(a) and 5(b)), and 10CFn showed an increasing evaluation for progressively worse solutions (Figure 9(a)). 11AARInm and 13AGRI identified the reference solution, attained their maxima 1 for the reference clustering, and were sensitive to the difference in the numbers of clusters in all scenarios.

Measures	EHC	FC^{FCM}	FC^{EMGM}	NEHC	PC
JI	$k^*/1$	-	-	-	-
RI	$k^*/1$	-	-	-	-
ARI	$k^*/1$	-	-	-	-
BC	$k^*/1$	-	-	-	-
03MI	$k^*/1$	$k^*/\cdot$	$k^*/\cdot$	-	-
05MI	$k^*/1$	$k^*/\cdot$	$k^*/\cdot$	-	-
03VI	$k^*/1$	$\cdot/\cdot$	$k^*/\cdot$	-	-
07CRI	$k^*/1$	$k^*/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$	$k^*/\cdot$
07CARI	$k^*/1$	$k^*/\cdot$	$k^*/\cdot$	$\cdot/\cdot$	$k^*/\cdot$
08BRIp	$k^*/1$	$\cdot/\cdot$	$\cdot/\cdot$	-	-
08BRIm	$k^*/1$	$k^*/\cdot$	$\cdot/\cdot$	-	-
09EBC	$k^*/1$	-	-	$k^*/1$	-
09CRI	$\cdot/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$
09CARI	$\cdot/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$
09HI	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$
09RI	$k^*/1$	$\cdot/\cdot$	$\cdot/\cdot$	-	-
09BRI	$k^*/1$	$k^*/\cdot$	$\cdot/\cdot$	$k^*/\cdot$	$k^*/\cdot$
09BARI	$k^*/1$	$k^*/\cdot$	$k^*/\cdot$	$k^*/\cdot$	$k^*/\cdot$
10QRIp	$k^*/1$	$k^*/\cdot$	$\cdot/\cdot$	$k^*/\cdot$	$k^*/\cdot$
10QRIm	$k^*/1$	$k^*/\cdot$	$\cdot/\cdot$	$k^*/\cdot$	$k^*/\cdot$
10ARI	$k^*/1$	$\cdot/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$
10AARI	$k^*/1$	$k^*/\cdot$	$k^*/\cdot$	$\cdot/\cdot$	$k^*/1$
10ARIn	$k^*/1$	$\cdot/\cdot$	$\cdot/\cdot$	$\cdot/\cdot$	$k^*/1$
10AARIn	$k^*/1$	$k^*/\cdot$	$k^*/\cdot$	$\cdot/\cdot$	$k^*/1$
10CSI	$k^*/1$	-	-	$k^*/1$	-
10CF	$k^*/1$	$k^*/\cdot$	$k^*/\cdot$	$k^*/1$	$k^*/\cdot$
10CFn	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$
11ARInm	$k^*/1$	$k^*/1$	$\cdot/\cdot$	$\cdot/\cdot$	$k^*/1$
11AARInm	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$
11MD	$k^*/1$	-	-	$k^*/1$	-
11D2	$k^*/1$	-	-	$k^*/1$	-
13GRI	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$
13AGRI	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$	$k^*/1$
12DB	$k^*/1$	-	-	$\cdot/\cdot$	-

"k^*" means that the measure identified the reference clustering, and "1" means that the measure attained its maximum 1 for the identified reference clustering. A cell with "-" denotes that the measure was not developed for the corresponding clustering type.

Table 3: Maximum and discriminant properties displayed by measures.

7.2 Comparing Randomly Generated Clusterings

The experiment in this section is based on a previously published one (Vinh et al., 2009, 2010) that assessed the ability of proposed EHC measures (based on information theory) to yield a constant baseline for randomly generated solutions. For a particular clustering type (EHC, FC, NEHC, or PC), random model (uniform, beta, unbalanced, or unbalanced-beta), 2-tuple (n, k^*), and $k \in \{2, 3, \ldots, 2k^*\}$, we generated 30 clustering pairs with n objects. Each pair contains a clustering with k clusters (representing an obtained solution) and a clustering with k^* clusters (representing a reference solution). We used four combinations of the number of objects and the true number of clusters: $(n = 25, k^* = 5)$, $(n = 100, k^* = 5)$, $(n = 50, k^* = 10)$, and $(n = 200, k^* = 10)$. The random models used to generate the clusterings depended on the clustering type as follows:

- For EHC, we generated clusterings for both the uniform and unbalanced models. In the uniform model, each object was uniformly assigned to one cluster. In the unbalanced model, each object was assigned to one cluster according to the following distribution: $p_1 \triangleq 0.1/k$ and $p_j \triangleq p_{j-1} + \alpha$ s.t. $\sum_{j=1}^{k} p_j = 1$ (it implies that $\alpha = 1.8/(k(k-1))$), where p_j is the probability of assigning an object to the jth cluster;

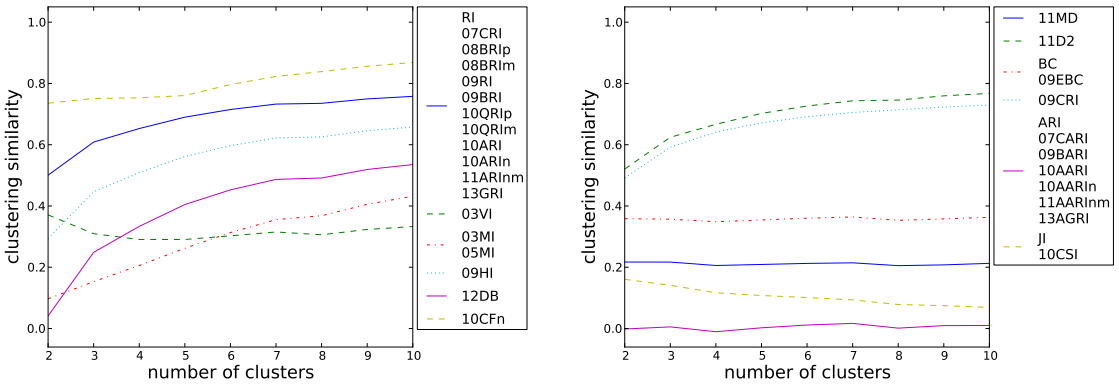
- For FC, we generated clusterings for the uniform, beta, and uniform-beta models. Let X_r^u be a random variable distributed according to the uniform distribution $\mathcal{U}(0, 1)$. For the uniform model, object x_i has a degree of membership to the rth cluster distributed according to $X_r^u/(X_1^u + X_2^u + \cdots + X_k^u)$, where k is the number of clusters. For the beta model, we uniformly draw $r_i \in \mathbb{N}_{1,k}$ for each object x_i to indicate to which cluster x_i probably has the highest degree of membership. Formally, let X_r^b and Y^b be two random variables distributed according to the beta distributions $\mathrm{Be}(1, 5)$ and $\mathrm{Be}(5, 1)$, respectively. Object x_i has a degree of membership to the rth cluster $(r \neq r_i)$ distributed according to $X_r^b/(X_1^b + \cdots + X_{r_i-1}^b + Y^b + X_{r_i+1}^b + \cdots + X_k^b)$ and to the r_ith cluster distributed according to $Y^b/(X_1^b + \cdots + X_{r_i-1}^b + Y^b + X_{r_i+1}^b + \cdots + X_k^b)$. The unbalanced-beta is equal to the beta model except that $r_i \triangleq 1$, such that the first cluster will have most of the membership;

- For NEHC, we generated clusterings for both the uniform and unbalanced models. In the uniform model, each object x_i was uniformly assigned to $k_i \in \mathbb{N}_{1,k}$ clusters, where k_i was uniformly drawn. In the unbalanced model, each object x_i was assigned to $k_i \in \mathbb{N}_{1,k}$ clusters according to the following method. Object x_i is assigned to a cluster according to the distribution p as in the EHC unbalanced model. The distribution p is then adjusted such that the cluster already drawn (say, the jth cluster) will not be selected again for x_i (i.e., $p_j \leftarrow 0$) and normalized to sum 1. The second cluster is randomly selected according to the resulting p. This process is repeated until x_i is assigned to k_i clusters;

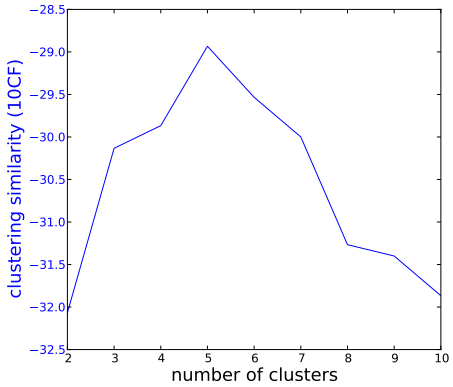
- For PC, we generated clusterings for the uniform, beta, and uniform-beta models. The distributions used are similar to those used for FC. The only difference is the absence of normalizing denominators in their definitions.

Cluster	(EHC, Un)	(FC, UBe)	(NEHC, Un)	(PC, UBe)
1st	2	10.2	22	15.9
2nd	11	10.4	53	16.4
3rd	20	11.4	64	17.0
4th	29	11.7	73	19.0
5th	38	56.2	74	83.3

Table 4: Object-to-cluster membership sums for clustering samples having $n = 100$ objects and $k^* = 5$ clusters.



(a) (b)



(c)

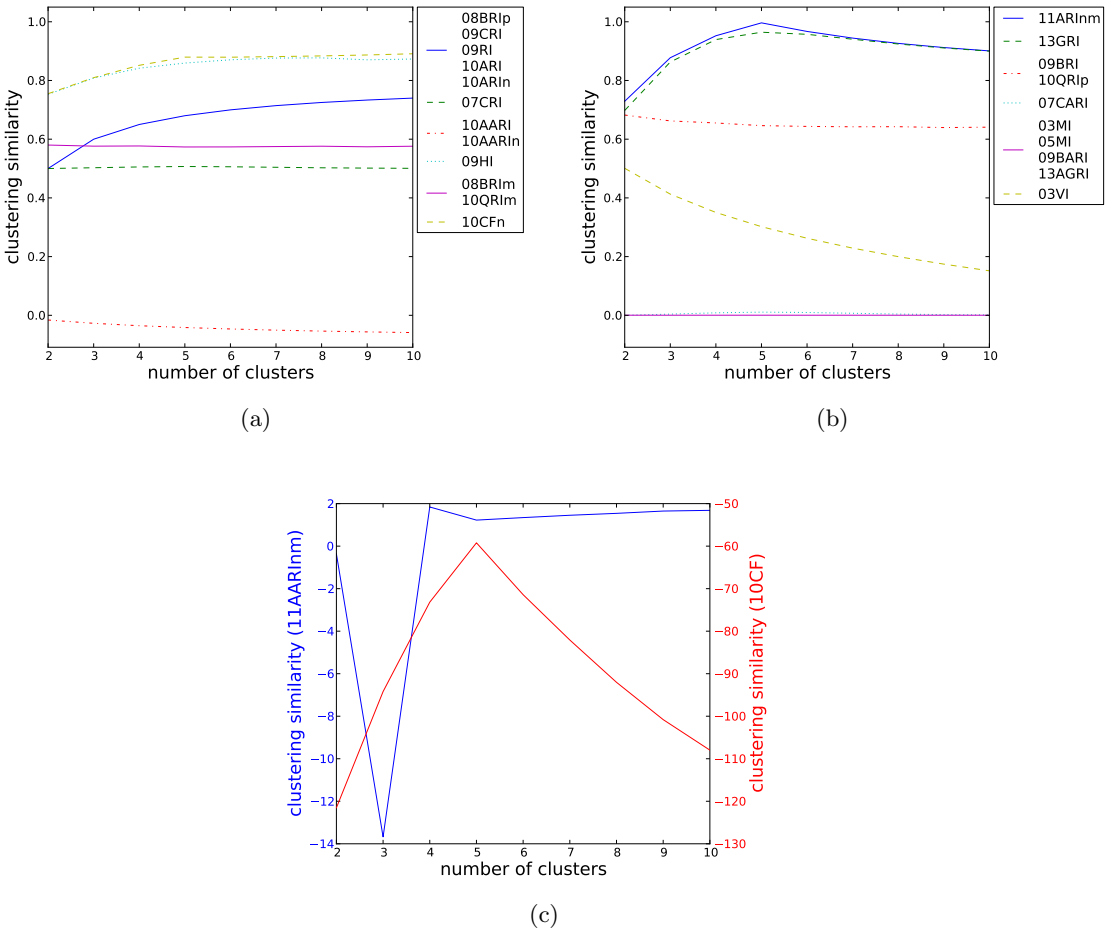
Figure 10: Average evaluations for $(\text{EHC}, \mathcal{U}, n = 25, k^* = 5)$.

We denote a particular experimental setting using a 4-tuple. For example, $(\text{EHC}, \mathcal{U}, n = 25, k^* = 5)$ refers to an EHC set generated according to the uniform model, where each clustering has 25 objects. The solutions of $(\text{EHC}, \mathcal{U}, n = 25, k^* = 5)$ were arranged in 30 EHC pairs for each $k \in \{2, 3, \ldots, 10\}$. Each pair contains an EHC with k clusters and an EHC with k^* clusters. Thus, the set $(\text{EHC}, \mathcal{U}, n = 25, k^* = 5)$ has $30 \cdot 9 = 270$ pairs of clusterings. The measures were then applied to evaluate the similarity between the two clusterings of each EHC pair, and the average evaluation for each $k \in \{2, 3, \ldots, 10\}$ was calculated and plotted in Figure 10. Similarly, Figures 11, 12, and 13 refer to the experimental settings $(\text{FC}, \mathcal{U}, n = 100, k^* = 5)$, $(\text{NEHC}, \mathcal{U}, n = 50, k^* = 10)$, and $(\text{PC}, \text{Be}, n = 200, k^* = 10)$, respectively. The remaining figures are not shown here to avoid cluttering but can be found in the supplementary material: `http://sn.im/25a9h8u`. Those figures will be referred here when appropriate.

Figures 10(a) and 10(b) show that 11 measures exhibited the same averages as RI and that six measures displayed the same averages as ARI, respectively. RI and JI (to a lesser extent) do not show a constant baseline (Hubert and Arabie, 1985; Albatineh et al., 2006), and this behavior is again observed in Figures 10(a) and 10(b). The 13GRI and 13AGRI measures showed the same averages as RI and ARI, respectively, because of their equivalence in the EHC context (Corollaries 4 and 5 in Appendix). 10CF attained a peak at $k = k^*$ clusters in Figure 10(c) for randomly generated clusterings. BC, 09EBC, 11MD, ARI, and the measures with similar values to ARI are the only ones that showed a constant baseline. The others showed a tendency to favor solutions with a high or low numbers of clusters.

Figure 11 shows the results for the experimental setting $(\text{FC}, \mathcal{U}, n = 100, k^* = 5)$. 03MI, 05MI, 07CARI, 09BARI, and 13AGRI displayed a constant baseline close to zero in Figure 11(b). 07CRI, 08BRIm, and 10QRIm (Figure 11(a)) also showed constant baselines, although not close to zero. These three measures were neither formally adjusted for chance nor based on a measure that was. Moreover, 07CRI, 08BRIm, and 10QRIm showed a low variance for a wide range of numbers of clusters in Figure 6. This leads us to suspect that the uniform behavior presented in Figure 11(a) is due to a poor sensitivity to solution variations. 09BRI and 10QRIp exhibited in Figure 11(b) a monotonically decreasing curve with low variation in values, as well as 10AARI and 10AARIn in Figure 11(a). 11AARInm produced values greater than its supposed maximum 1 and showed a counterintuitive behavior in Figure 11(c). 10CF, 11ARInm, and 13GRI showed a peak at $k = k^*$ for randomly generated clusterings.

07CARI, 09BARI, and 13AGRI are the only measures that displayed an approximately constant baseline close to zero in Figure 12, corresponding to the results for $(\text{NEHC}, \mathcal{U}, n = 50, k^* = 10)$. As for 10QRIm in Figure 11(a), the 10QRIp measure had a constant baseline in Figure 12(a) probably due to a low sensitivity in solution discrimination, as it is not adjusted for chance and is based on a measure (RI) known to be biased. The same cannot be said about 13AGRI, as it compares the solutions against a null model and exhibited a strong sensitivity in all experiments in Section 7.1. 10AARI showed in Figure 12(b) values greater than 1 for most solutions. 10CF (Figure 12(e)) and 13GRI (Figure 12(b)) again showed a peak at $k = k^*$ for randomly generated solutions. 10ARI and 11AARInm (Figure 12(d)) produced highly irregular evaluations. 11AARInm produced $-\infty$ (overflow) for $k = 2$ due to near-zero division.



(a) (b)

(c)

Figure 11: Average evaluations for $(\mathrm{FC}, \mathcal{U}, n = 100, k^* = 5)$.

Figure 13 illustrates the results for $(\mathrm{PC}, \mathrm{Be}, n = 200, k^* = 10)$. 09BRI, 09BARI, 10QRIp, 10QRIm, and 13AGRI showed constant baselines, and the constant baselines of 13AGRI and 09BARI were close to zero. 10CF (Figure 13(d)) and 13GRI (Figure 13(b)) again scored random clusterings with $k = k^*$ as better solutions. 10AARI and 11AARInm displayed highly unexpected values (Figure 13(c)).

Table 5 denotes which measures showed the baseline property. The italic n's refer to measures that provided constant baselines in the experiments corresponding to Figures 10, 11, 12, and 13 but not for all the remaining experiments. For example, BC and 09EBC showed unbiased evaluations in Figure 10(b) but not in the experiment $(\mathrm{EHC}, \mathrm{Un}, n = 100, k^* = 5)$ reported in the supplementary material.

Most measures could not provide an unbiased evaluation. They usually tend to favor random solutions with high or low numbers of clusters or show a peak in evaluating random solutions with the same number of clusters as the reference one. This behavior is undesirable, as the compared solutions were independently generated. Only 09BARI and 13AGRI

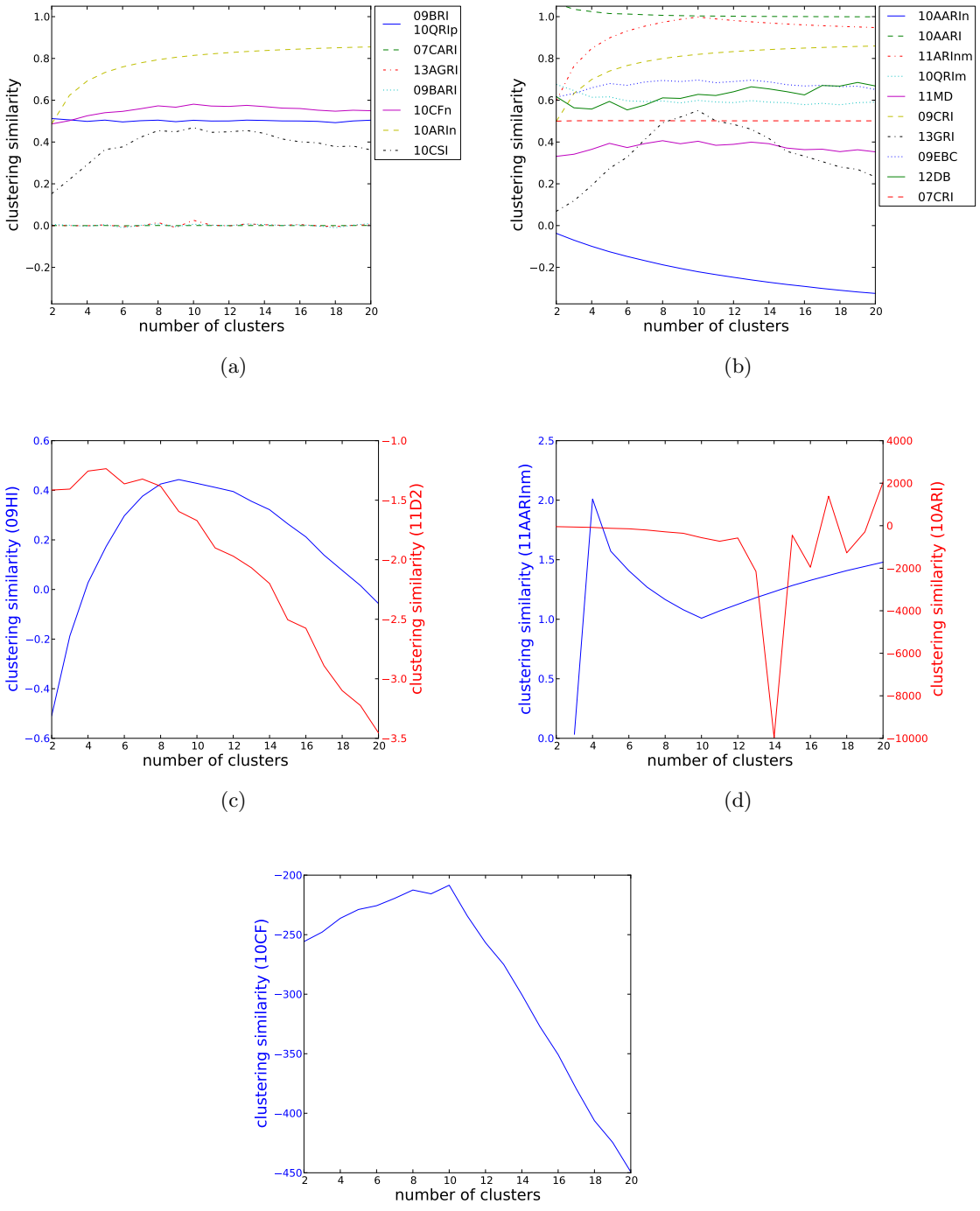


Figure 12: Average evaluations for (NEHC, $\mathcal{U}, n = 50, k^* = 10$).

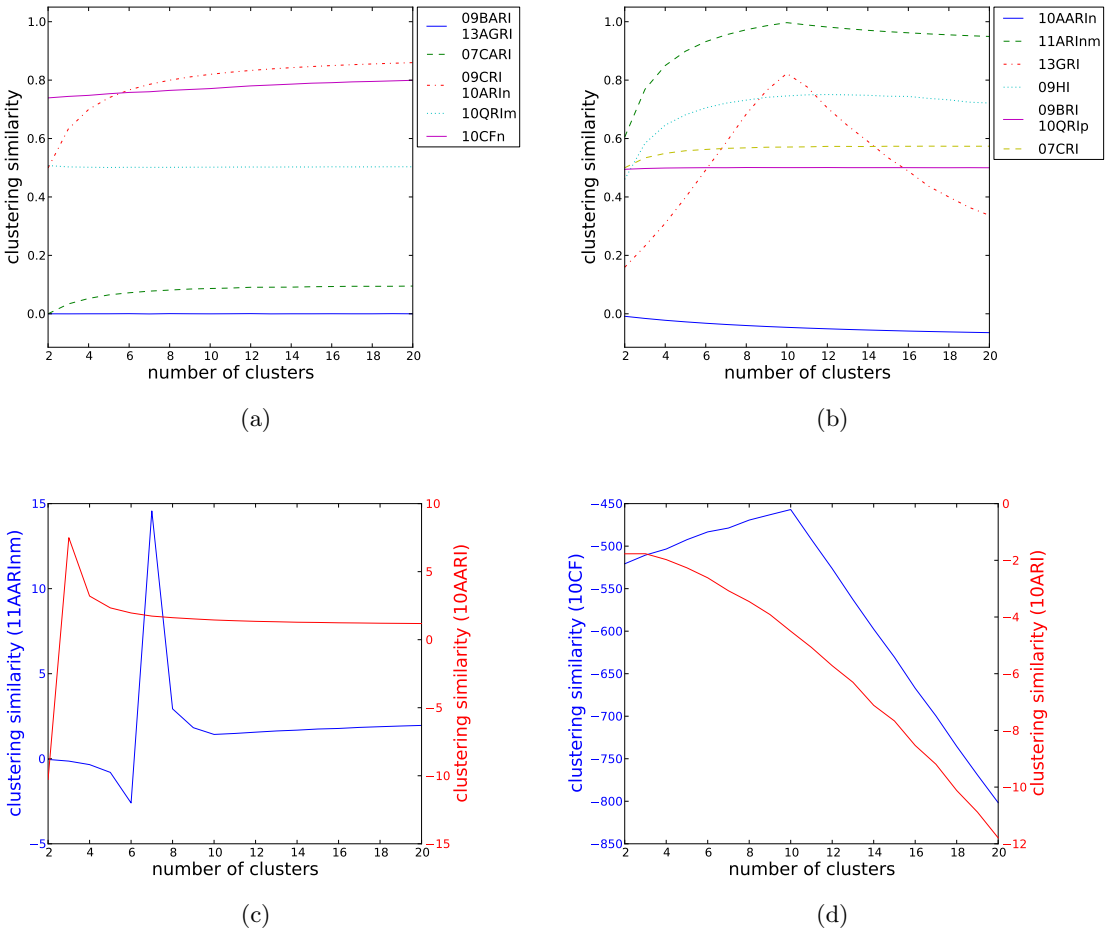


Figure 13: Average evaluations for $(\mathrm{PC}, \mathrm{Be}, n = 200, k^* = 10)$.

presented an approximately constant (and close to zero) baseline in all scenarios. The null model of 13AGRI is clearly violated in each scenario, which suggests that adjusting 13GRI is not just a theoretical adornment but a true correction that makes practical clustering comparisons fairer. Recall that, contrary to 13AGRI, 09BARI did not assign the maximum score 1 to the perfect solutions for all but the EHC scenario in the previous section.

7.3 13AGRI Evaluation Validity for FCs

We applied the k-means and FCM algorithms 30 times for each number of clusters $k \in \{2, 3, \ldots, 20\}$ to the UCI data sets (Newman and Asuncion, 2010) shown in Table 6. 13AGRI evaluated the best clustering (according to the respective algorithm's cost function) for each number of clusters using the known classification as the reference solution; the reference solution is thus an EHC. 13AGRI provides the same evaluation as ARI for k-means solutions since k-means produces EHCs (Corollary 5 in Appendix). FCM is regarded as the fuzzy version of k-means, both search for spherical-like clusters, and FCM tends to k-means when

Measures	EHC	FC	NEHC	PC	Measures	EHC	FC	NEHC	PC
JI	n	-	-	-	09BARI	**y**	**y**	**y**	**y**
RI	n	-	-	-	10QRIp	n	n	*n*	*n*
ARI	**y**	-	-	-	10QRIm	n	*n*	n	*n*
BC	*n*	-	-	-	10ARI	n	n	n	n
03MI	n	**y**	-	-	10AARI	**y**	n	n	n
05MI	n	**y**	-	-	10ARIn	n	n	n	n
03VI	n	n	-	-	10AARIn	**y**	n	n	n
07CRI	n	*n*	**y**	n	10CSI	n	-	n	-
07CARI	**y**	*n*	**y**	n	10CF	n	n	n	n
08BRIp	n	n	-	-	10CFn	n	n	n	n
08BRIm	n	*n*	-	-	11ARInm	n	n	n	n
09EBC	*n*	-	n	-	11AARInm	**y**	n	n	n
09CRI	n	n	n	n	11MD	*n*	-	n	-
09CARI	-	-	-	-	11D2	n	-	n	-
09HI	n	n	n	n	13GRI	n	n	n	n
09RI	n	n	-	-	13AGRI	**y**	**y**	**y**	**y**
09BRI	n	n	*n*	*n*	12DB	n	-	n	-

Table 5: Did the similarity measure display approximately constant baselines?

FCM exponent m approaches 1 (Yu et al., 2004). Thus, their solutions are often similar in the sense that converting an FCM solution into an EHC (by assigning the objects to the clusters for which they have the highest membership degrees) results in a clustering in which the relative assignment of objects is similar to the relative assignment of objects in the solution produced by k-means (i.e., when objects x_i and x_j are assigned to the same cluster in one solution, they are often assigned to the same cluster in the other solution). This section examines whether 13AGRI produces similar evaluations for solutions generated by k-means and FCM. If this is the case, we can be more confident in the validity of 13AGRI FC evaluations since 13AGRI and ARI are equivalent in the EHC domain.

For each data set, Table 7 displays the Pearson correlations between 13AGRI evaluations of the solutions produced by k-means and of the solutions produced by FCM across the number of clusters in $\{2, 3, \ldots, 20\}$. Five correlations were higher than 0.9, and more than a half were higher than 0.7. Figures 14(a) and 14(b) depict 13AGRI evaluations for the data sets on which the correlations attained the three highest and three lowest values, respectively. Figure legends display the corresponding data set, clustering type, and the number of classes in the a priori classification. Because the reference solutions are EHCs, 13AGRI almost always provided higher scores when evaluating EHC solutions than when evaluating FC solutions. The lowest correlations seem to have been obtained in the data sets for which the algorithms could not find good clusterings. For these data sets, the similarity between the found solutions and the reference one mostly fluctuates across the numbers of clusters as (we conjecture) there is no ideal number of clusters at which a peak on the

1. The original data set has 16 objects with missing attributes. We adopted the k-nearest neighbor algorithm with Euclidean distance for imputation (Hastie et al., 1999) and used the resulting data set.

Name	# Objects	# Attributes	# Classes
Breast cancer w. d. (bcw-d)	569	30	2
Breast cancer w. o. (bcw-o)[1]	699	9	2
Synthetic control chart (chart)	600	60	6
Ecoli data set (ecoli)	336	7	7
Glass identification (glass)	214	9	6
Haberman (haberman)	306	3	2
Image segmentation (img)	210	19	7
Ionosphere (ion)	351	34	2
Iris (iris)	150	4	3
Pima indians diabetes (pima)	768	8	2
Connectionist bench (sonar)	208	60	2
SPECT heart (heart)	267	22	2
Vehicle silhouettes (vehicle)	846	18	4
Wine (wine)	178	13	3

Table 6: UCI data sets.

evaluation curve would be found. K-means and FCM produced rather poor solutions for the haberman and sonar data sets according to 13AGRI. 13AGRI evaluations indicate that k-means could uncover some structure in the chart data set because a 13AGRI score (also an ARI score) of 0.5 is a considerable one according to our experience. However, there was not a distinctive solution across the numbers of clusters. 13AGRI indicates the FC solution with three clusters as the most similar to the reference one for the chart data set.

To further investigate the behavior of 13AGRI for the chart solutions, we reduced the chart dimensionality by projecting the 60-dimensional data to the first nine principal components (Jolliffe, 2002) explaining 90% of the variance. We identified two pairs of classes with high degree of overlap (namely, classes *decreasing trend* with *downward shift* and *increasing trend* with *upward shift* (Alcock, 1999)) by projecting the data onto several planes. We joined the classes *decreasing trend* with *downward shift* and *increasing trend* with *upward shift*, resulting in a classification (used as the reference clustering) with four classes. The Pearson correlation between 13AGRI evaluations is now 0.91 using the same experimental configuration as above. Figure 15 shows the evaluations for k-means and FCM solutions. 13AGRI provided high evaluations for k-means solutions with three and four clusters, while 13AGRI suggests that the best FCM solution is the one with three clusters.

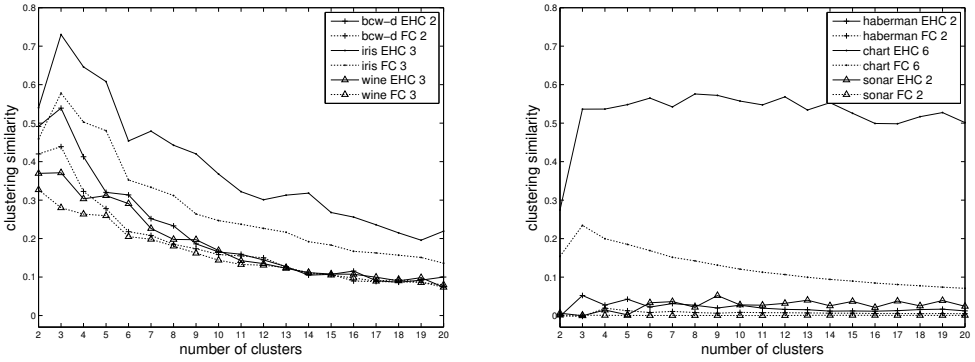
Results indicate that 13AGRI when applied to FCs behaves similarly to 13AGRI (i.e., ARI) when applied to EHCs, particularly when the solutions uncover some data set structure. Considering that ARI is one of the most trusted similarity measures, the results corroborate the 13AGRI evaluation validity for FCs.

7.4 Clustering Stability Assessment

We applied EMGM to subsamples of the top five data sets from the previous section (i.e., bcw-d, iris, wine, bcw-o, and img) 100 times for each number of clusters $k \in \{2, \ldots, 20\}$, generating 100 Gaussian mixtures for each number of clusters and data set; these Gaussian

bcw-d	iris	wine	bcw-o	img	ecoli	ion
0.99	0.99	0.98	0.98	0.91	0.89	0.83
vehicle	glass	pima	heart	haberman	chart	sonar
0.75	0.70	0.69	0.60	0.23	0.02	-0.45

Table 7: Correlation between 13AGRI evaluations of hard exclusive and fuzzy clusterings.



(a) (b)

Figure 14: 13AGRI evaluations that exhibited the three highest (a) and the three lowest correlations (b).

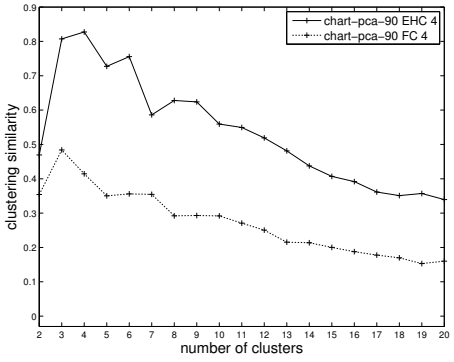


Figure 15: 13AGRI evaluations for the processed chart data set.

bcw-d	bcw-o	wine	iris	img
0.94	0.90	0.85	0.67	0.59

Table 8: Correlation between 13AGRI evaluation and stability statistic.

mixtures are different explanations for the phenomenon that produced the data set. We calculated a probabilistic clustering U (also known as FC) of the whole data set for each Gaussian mixture such that $U_{r,i}$ is the probability of x_i belonging to the rth cluster (i.e., to the rth Gaussian mixture component). 13AGRI compared each of the $\binom{100}{2}$ probabilistic clustering pairs for each number of clusters and data set, and the average was taken as the stability statistic (the less diverse the solution set, the higher the stability statistic) for the corresponding number of clusters and data set. Subsamples were generated by randomly selecting 80% of the data set objects, without replacement, as in (Monti et al., 2003). Algorithm 2 describes how stability assessment can be used to estimate the number of clusters and to select a promising clustering of a set of solutions.

Algorithm 2 Stability assessment

Require: Data set X.

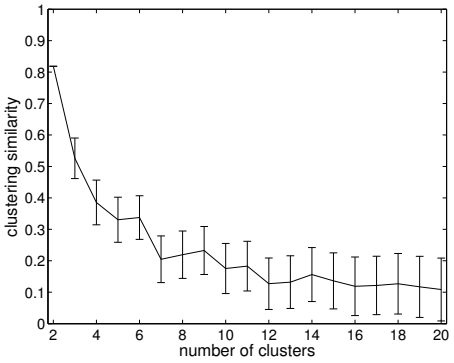
1: **for** $i \in \{1, 2, \ldots, 100\}$ **do**
2: $S_i \leftarrow$ Randomly draw 80% of the objects from X, without reposition.
3: **end for**
4: **for** $k \in \{2, 3, \ldots, 20\}$ **do**
5: **for** $i \in \{1, 2, \ldots, 100\}$ **do**
6: Apply EMGM to S_i finding a Gaussian mixture with k components.
7: $U^i \leftarrow$ Calculate the probabilistic clustering of the whole data set using the found Gaussian mixture.
8: **end for**
9: $t_k \leftarrow \sum_{i<j} 13\text{AGRI}(U^i, U^j)/\binom{100}{2}$ {stability statistic}
10: $V^k \leftarrow \text{argmax}_{U^i}\{\sum_{j \neq i} 13\text{AGRI}(U^i, U^j)\}$ {clustering set prototype}
11: **end for**
12: $k' \leftarrow \text{argmax}_{k \in \{2,\ldots,20\}}\{t_k\}$ {estimated number of clusters}
13: $U' \leftarrow V^{k'}$; {estimated best clustering}

Table 8 shows the Pearson correlations between stability statistic (defined by Step 9) values and 13AGRI evaluations (similarity between prototype V^k, Step 10, and the reference clustering) for different number of clusters. The high correlations indicate that the stability statistic, which can be used in real scenarios, approximately follows the 13AGRI evaluation that depends on a reference solution.

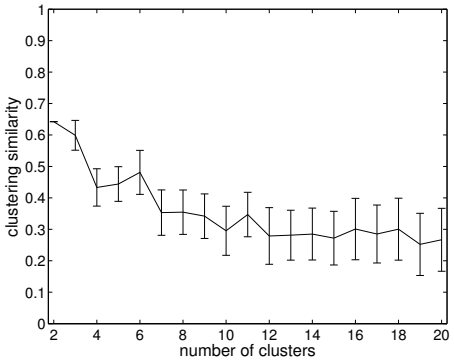
Figure 16 depicts 13AGRI evaluation for each clustering set prototype (Step 10 in Algorithm 2) and data set. We generated the error bar for a given $k \in \{2, \ldots, 20\}$ and data set as follows. Let t_k be the stability statistic for the set of clusterings with k clusters each (Step 9). Error bar was calculated to take 0 for the more stable clustering set (highest stability statistic) and 0.1 for the least stable clustering set, for visualization purposes. Thus, the

error bar value corresponding to the set of clusterings with k clusters is

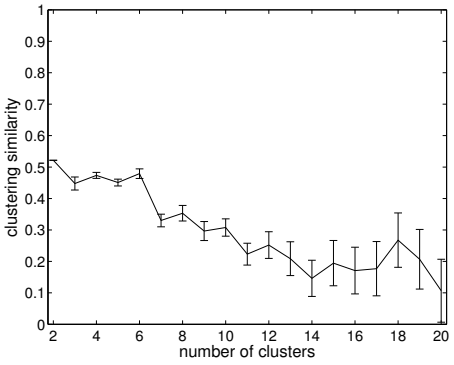
$$\frac{t_k - \min\{t_2, t_3, \ldots, t_{20}\}}{\max\{t_2, t_3, \ldots, t_{20}\} - \min\{t_2, t_3, \ldots, t_{20}\}} \times 0.1.$$



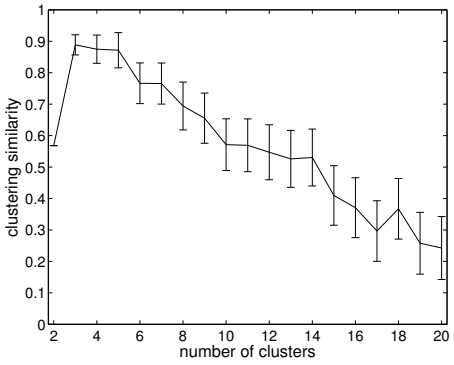
(a) Data set: bcw-d ($k^* = 2$).



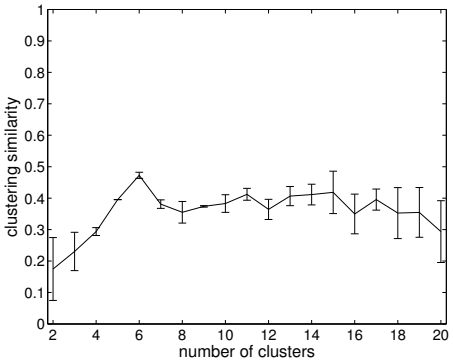
(b) Data set: bcw-o ($k^* = 2$).



(c) Data set: wine ($k^* = 3$).



(d) Data set: iris ($k^* = 3$).



(e) Data set: img ($k^* = 7$).

Figure 16: 13AGRI evaluations with error bars indicating clustering set instability.

Stability statistic precisely estimated the correct number of clusters for bcw-d (Figure 16(a)) and bcw-o (Figure 16(b)) data sets. The top two stable clustering sets in iris are the

ones with two and three clusters. Iris data set is classified in three classes (namely, *setosa*, *versicolour*, and *virginica*). However, it is well-known that the *versicolour* and *virginica* classes have a high degree of overlap and are frequently considered a single cluster (Wu and Yang, 2005), which corroborate the validity of the stability statistic. Although not being able to indicate the exact number of clusters, the lowest instability values for wine and img are around the correct number of clusters. In general, the near the number of clusters of the clustering set to the ideal one, the more stable the clustering set tends to be. These good preliminary results demonstrate that 13AGRI deserves further investigations on its applicability to the estimation of the number of clusters for FCs.

8. Discussion

Sections 7.1 and 7.2 empirically explored the four measure properties proposed in Section 2. Section 7.1 investigated the maximum, discriminant, and contrast properties by applying the measures to gradually different solutions. The hypothesis was that the similarity between the found clustering and the reference one is highly correlated to the difference in the number of clusters (epsilons in the case of SUBCLU) between the compared solutions, given that the solutions are produced by clustering algorithms capable of finding the ideal solution. One can understand the difference between the number of clusters given to the algorithm and the number of clusters of the reference solution as how far the domain of solutions of the corresponding algorithm is to the reference clustering. It is expected that a good measure should translate that difference in terms of evaluations. Section 7.1 showed that several of the measures did not follow the above hypothesis or did so in a very loose way, showing almost flat evaluations over the number of clusters. Moreover, several measures could not discriminate the best solution (03VI, 07CRI, 07CARI, 08BRIp, 08BRIm, 09CRI, 09CARI, 09RI, 09BRI, 10QRIp, 10QRIm, 10ARI, 10AARI, 10ARIn, 10AARIn, 11ARInm, and 12DB) for at least one of the clustering domains considered. We believe that this result by itself is enough for considering those measures unsuitable for the clustering domains they have failed. Section 7.1 concluded that 03MI, 05MI, 09BARI, and 10CF (beside the ones that have failed for the discriminant property) did not show the maximum property, and several measures were poorly sensitive to different solution qualities (poor contrast).

The baseline property was investigated in Section 7.2. In particular, we aimed to find out what measures were able to perform unbiased evaluations over different numbers of clusters. We concluded that only 09BARI and 13AGRI showed the baseline property for every clustering domain. By correcting 13GRI for chance, we were striving to build a measure that can capture the similarity between two solutions irrespectively to their numbers of clusters. We thus implicitly assumed that the number of clusters is not per se an indication of the similarity between clusterings (Section 7.2) but only a factor that delineates the domain of solutions (Section 7.1).

The correction-for-chance property implemented for 13AGRI, and that other measures displayed in Section 7.2 for certain scenarios, can also be understood as a way to stretch out the measure such that its useful range lies between the constant baseline and the maximum. As a matter of fact, one is not usually interested in very poor solutions (i.e., the ones that are far from the reference) (Meila, 2012), and those would receive negative or close to zero evaluations by 13AGRI and other adjusted measures. The correction-for-

chance thus increases the interpretability of the results by stressing what one should expect from clusterings whose evaluations lie below, close to, or above the baseline.

9. Conclusions

This paper discussed the importance of similarity measures in evaluating clustering algorithms, consensus clustering, clustering stability assessment, and quantifying information loss. These and other applications led to a recent interest in measures (especially pair-based ones) capable of comparing more general clusterings than the exclusive hard ones (usual partitions of an object set). We provided an overview of 28 measures proposed in the past 10 years for this task and discussed some of their issues. We showed that several of these measures do not attain the maximum whenever two equivalent solutions are compared and that most measures are biased toward clusterings with certain numbers of clusters. Moreover, several of the discussed measures are based on equations that were originally developed specifically for and by assuming the exclusive hard domain. Some measures thus exhibited unexpected behavior in experiments involving more general scenarios.

We proposed the 13FRI measure that can be used to compare fuzzy/probabilistic and exclusive hard clusterings. Based on a null model we proposed, according to which clusterings are generated, and following the framework employed by Hubert and Arabie (1985) to adjust the Rand index, 13AFRI was proposed as a corrected-for-chance version of 13FRI. We then extended 13FRI and 13AFRI to handle more general clusterings, namely possibilistic clusterings (including exclusive hard, fuzzy/probabilistic, and non-exclusive hard clusterings), yielding 13GRI and 13AGRI, respectively. The computational complexity analysis showed that our measures are practical.

In the first experiment involving four clustering algorithms of different natures, we observed that some measures could not identify the best solutions, and that several could not provide a fine-grained evaluation across the range of the numbers of clusters, whereas 13AGRI always attained its maximum 1 for the true number of clusters and displayed a steep, discriminative evaluation curve with a clear peak at the true number of clusters for each data set. We assessed the capability of the measures to provide an unbiased evaluation for randomly generated solutions with different numbers of clusters in the second experiment. A fair measure should assign a uniform evaluation across the range of the numbers of clusters, as each generated solution is independent of the reference one (Vinh et al., 2010). This is the case of the well-known adjusted Rand index (ARI) (Hubert and Arabie, 1985) for the exclusive hard domain. Only 13AGRI and 09BARI (Brouwer, 2009) (a recently proposed measure) displayed such an evaluation for all considered scenarios, which include the exclusive hard context; however, 09BARI could not attain its maximum 1 at the true number of clusters for all but the hard exclusive domain in the first experiment. The other measures exhibited a preference for certain solutions, which is attributable solely to their evaluation mechanisms. While the randomness model for 13AGRI incorporates some assumptions about the clusterings, those generated in our experiments clearly do not follow such a requirement. Even so, 13AGRI could provide uniform evaluations close to zero in the experiments with randomly generated solutions.

Two more experiments involving 14 real data sets and the algorithms k-means (MacQueen, 1967), fuzzy c-means (FCM) (Bezdek, 1981), and expectation maximization for

Gaussian mixtures (Dempster et al., 1977) were performed to assess the validity of 13AGRI evaluations in the fuzzy domain, arguably the most important domain after the exclusive hard one, and to investigate 13AGRI's applicability to the estimation of the number of clusters without (of course) any knowledge about the true data structure. We argue that the evaluations of the solutions produced by k-means and FCM for the same data set should be similar, and this behavior presented by 13AGRI is even more important for its validity because 13AGRI and the trusted ARI measures are equivalent when applied to solutions generated by k-means. The stability statistic based on 13AGRI defined in our last experiment showed good results indicating that 13AGRI can also be successfully applied to the estimation of the number of clusters in the probabilistic domain.

We proved that 13AGRI and ARI are equivalent in the exclusive hard domain. This is reassuring because (i) ARI is one of the most trusted similarity measures (Steinley, 2004; Albatineh et al., 2006), and (ii) the null model of 13AGRI was developed for general possibilistic clusterings (including exclusive hard clusterings as a special case). As future work, we think that 13AGRI deserves a further investigation on its conceptual properties, specially those generally taken as useful for similarity measures for clustering, such as cluster homogeneity sensibility, cluster completeness, and metric axioms compliance (Meila, 2007; Amigó et al., 2009).

Acknowledgments

We thank the editor and the anonymous reviewers for their constructive comments. This work was financially supported by the Brazilian agencies CNPq (#304137/2013-8) and FAPESP (#2009/17469-6 & #2013/18698-4).

Appendix A.

Proposition 1 *Let* U *and* V *be two FCs such that* $13\text{FRI}(\mathrm{U},\mathrm{V}) = 0$*,* $n > 1$*, and* $1 \leq k_{\mathrm{U}}, k_{\mathrm{V}} \leq n$*. It implies that* U *and* V *are EHCs and that* $k_{\mathrm{U}} = 1$ *and* $k_{\mathrm{V}} = n$ *or* $k_{\mathrm{U}} = n$ *and* $k_{\mathrm{V}} = 1$*, which unambiguously determine* U *and* V*.*

Proof Realize from Equations (12) that $\sum_{r=1}^{k_{\mathrm{U}}} \mathrm{U}_{r,l} = 1\ \forall l$ implies $\mathrm{S}^{\mathrm{U}}_{i,j} = 1 - \mathrm{J}^{\mathrm{U}}_{i,j}$. To have $13\text{FRI}(\mathrm{U},\mathrm{V}) = 0$, it must be the case that $\dot{a} = \dot{d} = 0$ (Equation 14), which implies that $\min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\} = \min\{1 - \mathrm{J}^{\mathrm{U}}_{i,j}, 1 - \mathrm{J}^{\mathrm{V}}_{i,j}\} = 0\ \forall i < j$ (Equations 13a and 13d). Hence, $\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j} \in \{0,1\}$ and $\mathrm{J}^{\mathrm{U}}_{i,j} \neq \mathrm{J}^{\mathrm{V}}_{i,j}$ for all $i < j$.

We first prove by contradiction that U cannot have a column i and a row r for which $\mathrm{U}_{r,i} \in (0,1)$ (the same holds for V). Assuming that the ith column of U has $\mathrm{U}_{r,i} \in (0,1)$ for an $r \in \mathbb{N}_{1,k_{\mathrm{U}}}$, we have $k_{\mathrm{U}} > 1$ and at least two elements of $\mathrm{U}_{:,i}$ have values in the open interval $(0,1)$ because $\sum_{t=1}^{k_{\mathrm{U}}} \mathrm{U}_{t,i} = 1$. Without loss of generality, assume that $i = 1$ (the columns of U and V can always be simultaneously permuted without changing the measure). We know that $\mathrm{U}^{\mathrm{T}}_{:,1}\mathrm{U}_{:,j} = \mathrm{J}^{\mathrm{U}}_{1,j} = 0\ \forall j \in \mathbb{N}_{2,n}$ because $\mathrm{U}^{\mathrm{T}}_{:,1}\mathrm{U}_{:,j}$ cannot yield 1. Thus, $\mathrm{J}^{\mathrm{V}}_{1,j} = 1\ \forall j \in \mathbb{N}_{2,n}$. This implies that the columns of V are all identical and each one has the element 1, resulting in $k_{\mathrm{V}} = 1$ because of the constraint $\sum_{j=1}^{n} \mathrm{V}_{t,j} > 0\ \forall t$. We thus have $\mathrm{J}^{\mathrm{V}}_{i_1,j_1} = 1\ \forall i_1 < j_1$ and $\mathrm{J}^{\mathrm{U}}_{i_2,j_2} = 0\ \forall i_2 < j_2$. The last equality only holds with constraint

$\sum_{j=1}^{n} \mathrm{U}_{t,j} > 0 \ \forall t$ if each row of U has exactly one value greater than zero. The property $\sum_{t=1}^{k_\mathrm{U}} \mathrm{U}_{t,j} = 1 \ \forall j$ of FCs and the assumption $k_\mathrm{U} \leq n$ then require each column of U to have exactly one value greater than zero (and to have $k_\mathrm{U} = n$ rows), which is the value 1. This violates the assumption that $\mathrm{U}_{r,i} \in (0,1)$, which implies that U (and V) must be a matrix with only zeros and ones.

Suppose $n = 2$. If columns 1 and 2 of U are identical, columns 1 and 2 of V are different because we have already proven that $\mathrm{J}^\mathrm{U}_{i,j} \neq \mathrm{J}^\mathrm{V}_{i,j}$. This only can happen for $k_\mathrm{U} = 1$ and $k_\mathrm{V} = 2$ (remember the properties of an FC). Now, suppose that $n > 2$ and, without loss of generality, that $\mathrm{U}_{:,1}$ and $\mathrm{U}_{:,2}$ are identical and that $\mathrm{V}_{:,1}$ and $\mathrm{V}_{:,2}$ are different. If a column $i > 2$ of U differs from columns 1 and 2 of U, we conclude that columns 1 and 2 of V are equal to column i of V. However, this implies that columns 1 and 2 of V are equal, and, as we known, they are not. Consequently, all columns of U must be identical and all columns of V must be different. This can only happen for $k_\mathrm{U} = 1$ and $k_\mathrm{V} = n$, which proves the proposition. ■

Proposition 2 *Given two EHCs* U *and* V*, we have* $13\mathrm{FRI}(\mathrm{U},\mathrm{V}) = \mathrm{RI}(\mathrm{U},\mathrm{V})$.

Proof Realize that $\dot{a}$, $\dot{b}$, $\dot{c}$, and $\dot{d}$ (Equations 13) are equivalent to a, b, c, and d (Equations 7) by assigning the values 0 and 1 to $\mathrm{J}^\mathrm{U}_{i,j}$ and $\mathrm{J}^\mathrm{V}_{i,j}$. ■

Proposition 3 *Given two EHCs* U *and* V*, we have* $13\mathrm{AFRI}(\mathrm{U},\mathrm{V}) = \mathrm{ARI}(\mathrm{U},\mathrm{V})$.

Proof Both ARI and 13AFRI use the framework of Equation (15). The expectation of ARI given U and V is $\mathrm{E}[\mathrm{ARI}]_{\mathrm{U},\mathrm{V}} = (\mathrm{E}[a]_{\mathrm{U},\mathrm{V}} + \mathrm{E}[d]_{\mathrm{U},\mathrm{V}})/(a+b+c+d)$ (Hubert and Arabie, 1985). We must therefore only show that $\mathrm{E}[a]_{\mathrm{U},\mathrm{V}} = \mathrm{E}[\dot{a}]_{\mathrm{U},\mathrm{V}}$ and $\mathrm{E}[d]_{\mathrm{U},\mathrm{V}} = \mathrm{E}[\dot{d}]_{\mathrm{U},\mathrm{V}}$, since $a = \dot{a}$, $b = \dot{b}$, $c = \dot{c}$, and $d = \dot{d}$ by Proposition 2. Let $\mathrm{J}^\mathrm{U} = \mathrm{U}^\mathrm{T}\mathrm{U}$, $\mathrm{J}^\mathrm{V} = \mathrm{V}^\mathrm{T}\mathrm{V}$, and $\mathrm{N} = \mathrm{U}\mathrm{V}^\mathrm{T}$. Because U and V are EHCs, we can rewrite $\min\{\mathrm{J}^\mathrm{U}_{i,j}, \mathrm{J}^\mathrm{V}_{i,j}\} = \mathrm{J}^\mathrm{U}_{i,j}\mathrm{J}^\mathrm{V}_{i,j}$. Both $\sum_{i<j} \mathrm{J}^\mathrm{U}_{i,j}$ and $\sum_{r=1}^{k_\mathrm{U}} \binom{\mathrm{N}_{r,+}}{2}$ count the number of unordered object pairs in the same cluster in U. We thus have

$$
\begin{aligned}
\mathrm{E}[\dot{a}]_{\mathrm{U},\mathrm{V}} &= \frac{2}{n(n-1)} \sum_{i_1<j_1} \mathrm{J}^\mathrm{U}_{i_1,j_1} \sum_{i_2<j_2} \mathrm{J}^\mathrm{V}_{i_2,j_2} \\
&= \sum_{r=1}^{k_\mathrm{U}} \binom{\mathrm{N}_{r,+}}{2} \sum_{t=1}^{k_\mathrm{V}} \binom{\mathrm{N}_{+,t}}{2} \Big/ \binom{n}{2} \\
&= \mathrm{E}[a]_{\mathrm{U},\mathrm{V}} \text{ (Equation (2) in (Hubert and Arabie, 1985)).}
\end{aligned}
$$

Because $\mathrm{J}^{\mathrm{U}}_{i,j} = 1 - \mathrm{S}^{\mathrm{U}}_{i,j}$ for EHCs, we have

$$\begin{aligned}
\mathrm{E}[\dot{d}]_{\mathrm{U,V}} &= \frac{2}{n(n-1)} \sum_{i_1<j_1} \sum_{i_2<j_2} (1 - \mathrm{J}^{\mathrm{U}}_{i_1,j_1})(1 - \mathrm{J}^{\mathrm{V}}_{i_2,j_2}) \\
&= \binom{n}{2} - \sum_{i_1<j_1} \mathrm{J}^{\mathrm{U}}_{i_1,j_1} - \sum_{i_2<j_2} \mathrm{J}^{\mathrm{V}}_{i_2,j_2} \\
&+ \sum_{i_1<j_1} \sum_{i_2<j_2} \mathrm{J}^{\mathrm{U}}_{i_1,j_1} \mathrm{J}^{\mathrm{V}}_{i_2,j_2} \Big/ \binom{n}{2} \\
&= \binom{n}{2} - \sum_{r=1}^{k_{\mathrm{U}}} \binom{\mathrm{N}_{r,+}}{2} - \sum_{t=1}^{k_{\mathrm{V}}} \binom{\mathrm{N}_{+,t}}{2} \\
&+ \sum_{r=1}^{k_{\mathrm{U}}} \binom{\mathrm{N}_{r,+}}{2} \sum_{t=1}^{k_{\mathrm{V}}} \binom{\mathrm{N}_{+,t}}{2} \Big/ \binom{n}{2} \\
&= \mathrm{E}[d]_{\mathrm{U,V}} \text{ (Equation (3) in (Hubert and Arabie, 1985) multiplied by } \binom{n}{2} \text{ and} \\
&\text{then subtracted by } \mathrm{E}[a]_{\mathrm{U,V}}).
\end{aligned}$$

■

Proposition 4 *Given two PCs* U *and* V*, we have* $\dot{a} + \dot{b} + \dot{c} + \dot{d} = \sum_{i<j} \min\{\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j}\}$.

Proof Let

$$\begin{aligned}
\dot{a}_{i,j} &\triangleq \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}, \\
\dot{b}_{i,j} &\triangleq \min\{\mathrm{J}^{\mathrm{U}}_{i,j} - \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}, \mathrm{S}^{\mathrm{V}}_{i,j} - \min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\}\}, \\
\dot{c}_{i,j} &\triangleq \min\{\mathrm{J}^{\mathrm{V}}_{i,j} - \min\{\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j}\}, \mathrm{S}^{\mathrm{U}}_{i,j} - \min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\}\}, \text{ and} \\
\dot{d}_{i,j} &\triangleq \min\{\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j}\}.
\end{aligned}$$

We prove the proposition by showing that

$$\dot{a}_{i,j} + \dot{b}_{i,j} + \dot{c}_{i,j} + \dot{d}_{i,j} = \min\{\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j}\}. \tag{25}$$

Table 9 shows the six rank combinations between the values of the pairs $(\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j})$, $(\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j})$, and $(\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j})$, covering all possible scenarios. Equation (25) is true for each scenario. For conciseness, let us show the proof for Combinations 1 and 3 only.

Assuming Combination 1, we have $\dot{a}_{i,j} = \mathrm{J}^{\mathrm{V}}_{i,j}$, $\dot{b}_{i,j} = 0$, $\dot{c}_{i,j} = 0$, and $\dot{d}_{i,j} = \mathrm{S}^{\mathrm{V}}_{i,j}$, and Equation (25) is true. Assuming Combination 3, we have $\dot{a}_{i,j} = \mathrm{J}^{\mathrm{V}}_{i,j}$, $\dot{b}_{i,j} = \min\{\mathrm{J}^{\mathrm{U}}_{i,j} - \mathrm{J}^{\mathrm{V}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j} - \mathrm{S}^{\mathrm{U}}_{i,j}\}$, $\dot{c}_{i,j} = 0$, and $\dot{d}_{i,j} = \mathrm{S}^{\mathrm{U}}_{i,j}$. Note that $\mathrm{T}^{\mathrm{U}}_{i,j} < \mathrm{T}^{\mathrm{V}}_{i,j} \Rightarrow \mathrm{J}^{\mathrm{U}}_{i,j} + \mathrm{S}^{\mathrm{U}}_{i,j} < \mathrm{J}^{\mathrm{V}}_{i,j} + \mathrm{S}^{\mathrm{V}}_{i,j} \Rightarrow \mathrm{J}^{\mathrm{U}}_{i,j} - \mathrm{J}^{\mathrm{V}}_{i,j} < \mathrm{S}^{\mathrm{V}}_{i,j} - \mathrm{S}^{\mathrm{U}}_{i,j}$. Thus, $\dot{b}_{i,j} = \mathrm{J}^{\mathrm{U}}_{i,j} - \mathrm{J}^{\mathrm{V}}_{i,j}$, and Equation (25) is true. ■

#	$(\mathrm{J}^{\mathrm{U}}_{i,j}, \mathrm{J}^{\mathrm{V}}_{i,j})$	$(\mathrm{S}^{\mathrm{U}}_{i,j}, \mathrm{S}^{\mathrm{V}}_{i,j})$	$(\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j})$
1	$\mathrm{J}^{\mathrm{U}}_{i,j} \geq \mathrm{J}^{\mathrm{V}}_{i,j}$	$\mathrm{S}^{\mathrm{U}}_{i,j} \geq \mathrm{S}^{\mathrm{V}}_{i,j}$	$\mathrm{T}^{\mathrm{U}}_{i,j} \geq \mathrm{T}^{\mathrm{V}}_{i,j}$
2	$\mathrm{J}^{\mathrm{U}}_{i,j} \geq \mathrm{J}^{\mathrm{V}}_{i,j}$	$\mathrm{S}^{\mathrm{U}}_{i,j} < \mathrm{S}^{\mathrm{V}}_{i,j}$	$\mathrm{T}^{\mathrm{U}}_{i,j} \geq \mathrm{T}^{\mathrm{V}}_{i,j}$
3	$\mathrm{J}^{\mathrm{U}}_{i,j} \geq \mathrm{J}^{\mathrm{V}}_{i,j}$	$\mathrm{S}^{\mathrm{U}}_{i,j} < \mathrm{S}^{\mathrm{V}}_{i,j}$	$\mathrm{T}^{\mathrm{U}}_{i,j} < \mathrm{T}^{\mathrm{V}}_{i,j}$
4	$\mathrm{J}^{\mathrm{U}}_{i,j} < \mathrm{J}^{\mathrm{V}}_{i,j}$	$\mathrm{S}^{\mathrm{U}}_{i,j} \geq \mathrm{S}^{\mathrm{V}}_{i,j}$	$\mathrm{T}^{\mathrm{U}}_{i,j} \geq \mathrm{T}^{\mathrm{V}}_{i,j}$
5	$\mathrm{J}^{\mathrm{U}}_{i,j} < \mathrm{J}^{\mathrm{V}}_{i,j}$	$\mathrm{S}^{\mathrm{U}}_{i,j} \geq \mathrm{S}^{\mathrm{V}}_{i,j}$	$\mathrm{T}^{\mathrm{U}}_{i,j} < \mathrm{T}^{\mathrm{V}}_{i,j}$
6	$\mathrm{J}^{\mathrm{U}}_{i,j} < \mathrm{J}^{\mathrm{V}}_{i,j}$	$\mathrm{S}^{\mathrm{U}}_{i,j} < \mathrm{S}^{\mathrm{V}}_{i,j}$	$\mathrm{T}^{\mathrm{U}}_{i,j} < \mathrm{T}^{\mathrm{V}}_{i,j}$

Table 9: Rank combinations.

Corollary 1 *If* U *and* V *are two FCs with* n *columns each, we have* $\mathrm{T}^{\mathrm{U}}_{i,j} = \mathrm{T}^{\mathrm{V}}_{i,j} = 1$ *and the sum* $\dot{a} + \dot{b} + \dot{c} + \dot{d} = n(n-1)/2$.

Proposition 5 *Given two PCs* U *and* V*, we have* $\dot{a} + \dot{b} + \dot{c} + \dot{d} + \dot{e} = \max\{\sum_{i<j} \mathrm{T}^{\mathrm{U}}_{i,j}, \sum_{i<j} \mathrm{T}^{\mathrm{V}}_{i,j}\}$.

Proof Let $\mathrm{M} \triangleq \max\{\mathrm{T}^{\mathrm{U}}, \mathrm{T}^{\mathrm{V}}\}$. If $\mathrm{T}^{\mathrm{U}}_{i,j} \geq \mathrm{T}^{\mathrm{V}}_{i,j}$, then $\min\{\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j}\} + \mathrm{M}_{i,j} - \mathrm{T}^{\mathrm{V}}_{i,j} = \mathrm{T}^{\mathrm{U}}_{i,j}$. If $\mathrm{T}^{\mathrm{U}}_{i,j} < \mathrm{T}^{\mathrm{V}}_{i,j}$, then $\min\{\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j}\} + \mathrm{M}_{i,j} - \mathrm{T}^{\mathrm{V}}_{i,j} = \mathrm{T}^{\mathrm{U}}_{i,j}$ as well. Thus, $\mathrm{T}^{\mathrm{U}}_{i,j} = \min\{\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j}\} + \mathrm{M}_{i,j} - \mathrm{T}^{\mathrm{V}}_{i,j}$, and the same reasoning works for $\mathrm{T}^{\mathrm{V}}_{i,j} = \min\{\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j}\} + \mathrm{M}_{i,j} - \mathrm{T}^{\mathrm{U}}_{i,j}$. We know that $\dot{a} + \dot{b} + \dot{c} + \dot{d} = \sum_{i<j} \min\{\mathrm{T}^{\mathrm{U}}_{i,j}, \mathrm{T}^{\mathrm{V}}_{i,j}\}$ by Proposition 4. If $\sum_{i<j} \mathrm{T}^{\mathrm{U}}_{i,j} \geq \sum_{i<j} \mathrm{T}^{\mathrm{V}}_{i,j}$, we have $\dot{e} = \sum_{i<j} (\mathrm{M}_{i,j} - \mathrm{T}^{\mathrm{V}}_{i,j})$ and $\dot{a} + \dot{b} + \dot{c} + \dot{d} + \dot{e} = \sum_{i<j} \mathrm{T}^{\mathrm{U}}_{i,j}$; otherwise, $\dot{a} + \dot{b} + \dot{c} + \dot{d} + \dot{e} = \sum_{i<j} \mathrm{T}^{\mathrm{V}}_{i,j}$. ■

Corollary 2 *The sum* $\dot{a} + \dot{b} + \dot{c} + \dot{d} + \dot{e}$ *is constant over all simultaneous permutations of the columns of* U *and* V *because they do not alter the sums* $\sum_{i<j} \mathrm{T}^{\mathrm{U}}_{i,j}$ *and* $\sum_{i<j} \mathrm{T}^{\mathrm{V}}_{i,j}$.

Corollary 3 *13FRI (13AFRI) and 13GRI (13AGRI) are equivalent when applied to FCs because* $\max\{\sum_{i<j} \mathrm{T}^{\mathrm{U}}_{i,j}, \sum_{i<j} \mathrm{T}^{\mathrm{V}}_{i,j}\} = n(n-1)/2 = \dot{a} + \dot{b} + \dot{c} + \dot{d}$.

Corollary 4 *Given two EHCs* U *and* V*, we have* $\mathrm{13GRI}(\mathrm{U}, \mathrm{V}) = \mathrm{RI}(\mathrm{U}, \mathrm{V})$ *because of Proposition 2 and Corollary 3.*

Corollary 5 *Given two EHCs* U *and* V*, we have* $\mathrm{13AGRI}(\mathrm{U}, \mathrm{V}) = \mathrm{ARI}(\mathrm{U}, \mathrm{V})$ *because of Proposition 3 and Corollary 3.*

References

Ahmed N. Albatineh, Magdalena Niewiadomska-Bugaj, and Daniel Mihalko. On similarity indices and correction for chance agreement. *Journal of Classification*, 23:301–313, 2006. 10.1007/s00357-006-0017-z.

Robert Alcock. Synthetic control chart time series data set, 1999. URL `http://archive.ics.uci.edu/ml/datasets/Synthetic+Control+Chart+Time+Series`.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486, August 2009. ISSN 1386-4564.

Derek T. Anderson, James C. Bezdek, Mihail Popescu, and James M. Keller. Comparing fuzzy, probabilistic, and possibilistic partitions. *IEEE Trans. Fuzzy Syst.*, 18(5):906–918, June 2010.

Derek T. Anderson, James M. Keller, Ozy Sjahputera, James C. Bezdek, and Mihail Popescu. Comparing soft clusters and partitions. In *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, pages 924 –931, june 2011.

Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 79–85, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

Jrgen Beringer and Eyke Hllermeier. *Fuzzy Clustering of Parallel Data Streams*, pages 333–352. John Wiley & Sons, Ltd, 2007.

James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 1981. ISBN 0306406713.

Christian Borgelt. Resampling for fuzzy clustering. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, pages 595–614, 2007.

Christian Borgelt and Rudolf Kruse. Finding the number of fuzzy clusters by resampling. In *Fuzzy Systems, 2006 IEEE International Conference on*, pages 48–54, 0-0 2006.

Roelof Brouwer. Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. *Journal of Intelligent Information Systems*, 32:213–235, 2009. 10.1007/s10844-008-0054-7.

Ricardo 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):833 – 841, 2007.

Ricardo J. G. B. Campello. Generalized external indexes for comparing data partitions with overlapping categories. *Pattern Recognition Letters*, 31(9):966–975, 2010. ISSN 0167-8655.

Claire Cardie and Kiri Wagstaf. Noun phrase coreference as clustering. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC1999)*, pages 82–89, College Park, Maryland, USA, June 21–22 1999.

Michele Ceccarelli and Antonio Maratea. A fuzzy extension of some classical concordance measures and an efficient algorithm for their computation. In Ignac Lovrek, Robert

Howlett, and Lakhmi Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 5179 of *Lecture Notes in Computer Science*, pages 755–763. Springer Berlin / Heidelberg, 2008.

Michele Ceccarelli and Antonio Maratea. Concordance indices for comparing fuzzy, possibilistic, rough and grey partitions. *Int. J. Knowl. Eng. Soft Data Paradigm.*, 1(4): 331–344, October 2009.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246.

Brian S. Everitt, Sabine Landau, and Morven Leese. *Cluster Analysis*. Arnold Publishers, May 2001.

E. B. Fowlkes and C. L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

Ana L. N. Fred and Anil K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):835–850, 2005. ISSN 0162-8828.

Stephan Günnemann, Ines Färber, Emmanuel Müller, Ira Assent, and Thomas Seidl. External evaluation measures for subspace clustering. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1363–1372, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0717-8.

Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. Imputing missing data for gene expression arrays. Technical report, Stanford University, 1999.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2 (1):193–218, December 1985.

Eyke Hullermeier and Maria Rifqi. A fuzzy variant of the rand index for comparing clustering structures. In *Proc. IFSA*, page 16, Lisbon, Portugal, 2009.

Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Socit Vaudoise de Sciences Naturelles*, 44:223–370, 1908.

Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988. ISBN 0-13-022278-X.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, second edition, October 2002.

Karin Kailing, Hans-Peter Kriegel, and Peer Krger. Density-connected subspace clustering for high-dimensional data. In *Proceedings SDM (2004)*, pages 246–257, 2004.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley, 1990.

Ludmila I. Kuncheva and Dmitry P. Vetrov. Evaluation of stability of k-means cluster ensembles with respect to random initialization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1798 –1808, nov. 2006.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Ping Luo, Hui Xiong, Guoxing Zhan, Junjie Wu, and Zhongzhi Shi. Information-theoretic distance measures for clustering validation: Generalization and normalization. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1249–1262, September 2009.

David J. C. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 1 edition, June 2003. ISBN 0521642981.

James B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

Sara C. Madeira and Arlindo L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1):24–45, 2004.

Marina Meila. Comparing clusterings by the variation of information. In Bernhard Schlkopf and Manfred Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 173–187. Springer Berlin / Heidelberg, 2003. ISBN 978-3-540-40720-1.

Marina Meila. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pages 577–584, New York, NY, USA, 2005. ACM.

Marina Meila. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98:873–895, May 2007.

Marina Meila. Local equivalences of distances between clusterings–a geometric perspective. *Machine Learning*, 86(3):369–389, March 2012.

Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52:91–118, July 2003.

David Newman and Arthur Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

Malay K. Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy Sets and Systems*, 155(2):191 – 214, 2005. ISSN 0165-0114.

Nikhil R. Pal and James C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Trans. on Fuzzy Systems*, 3(3):370–379, 1995.

Anne Patrikainen and Marina Meila. Comparing subspace clusterings. *IEEE Trans. on Knowl. and Data Eng.*, 18(7):902–916, 2006.

Romain Quere and Carl Frelicot. A normalized soft window-based similarity measure to extend the rand index. In *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, pages 2513 –2520, june 2011.

Romain Quere, Hoel Le Capitaine, Noel Fraisseix, and Carl Frelicot. On normalizing fuzzy coincidence matrices to compare fuzzy and/or possibilistic partitions with the rand index. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pages 977–982, Washington, DC, USA, 2010. IEEE Computer Society.

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. ISSN 01621459.

Stefano Rovetta and Francesco Masulli. An experimental validation of some indexes of fuzzy clustering similarity. In *Proceedings of the 8th International Workshop on Fuzzy Logic and Applications*, WILF '09, pages 132–139, Berlin, Heidelberg, 2009. Springer-Verlag.

Douglas Steinley. Properties of the hubert-arabie adjusted rand index. *Psychological Methods*, 9(3):386–396, 2004.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, March 2003.

Nguyen X. Vinh and Julien Epps. A novel approach for automatic number of clusters detection in microarray data based on consensus clustering. In *Proceedings of the 2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering*, BIBE '09, pages 84–91, Washington, DC, USA, 2009. IEEE Computer Society.

Nguyen X. Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1073–1080, New York, NY, USA, 2009. ACM.

Nguyen X. Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

Zhimin Wang. Metrics for overlapping clustering comparison, November 2010. URL `http://etaxonomy.org/mw/File:Sigs.pdf`.

Zhimin Wang. Entropy on covers. *Data Mining and Knowledge Discovery*, 24:288–309, 2012. ISSN 1384-5810. 10.1007/s10618-011-0230-1.

Junjie Wu, Hui Xiong, and Jian Chen. Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 877–886, New York, NY, USA, 2009. ACM.

Kuo-Lung Wu and Miin-Shen Yang. A cluster validity index for fuzzy clustering. *Pattern Recognition Letters*, 26:1275–1291, July 2005. ISSN 0167-8655.

Jian Yu, Qiansheng Cheng, and Houkuan Huang. Analysis of the weighting exponent in the fcm. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(1):634 – 639, feb. 2004.

Zhiwen Yu, Hau-San Wong, and Hongqiang Wang. Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics*, 23(21):2888–2896, 2007.

Jiang-She Zhang and Yiu-Wing Leung. Improved possibilistic c-means clustering algorithms. *Fuzzy Systems, IEEE Transactions on*, 12(2):209–217, april 2004.

Shaohong Zhang, Hau-San Wong, and Ying Shen. Generalized adjusted rand indices for cluster ensembles. *Pattern Recognition*, 45(6):2214 – 2226, 2012.