Journal of Machine Learning Research 11 (2010) 2837-2854 Submitted 6/09; Revised 3/10; Published 10/10

Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance

Nguyen Xuan Vinh[*] N.X.VINH@UNSW.EDU.AU
Julien Epps[*] J.EPPS@UNSW.EDU.AU
The University of New South Wales
Sydney, NSW 2052, Australia

James Bailey[†] JBAILEY@CSSE.UNIMELB.EDU.AU
The University of Melbourne
Vic. 3010, Australia

Editor: Marina Meilă

Abstract

Information theoretic measures form a fundamental class of measures for comparing clusterings, and have recently received increasing interest. Nevertheless, a number of questions concerning their properties and inter-relationships remain unresolved. In this paper, we perform an organized study of information theoretic measures for clustering comparison, including several existing popular measures in the literature, as well as some newly proposed ones. We discuss and prove their important properties, such as the metric property and the normalization property. We then highlight to the clustering community the importance of correcting information theoretic measures for chance, especially when the data size is small compared to the number of clusters present therein. Of the available information theoretic based measures, we advocate the normalized information distance (NID) as a general measure of choice, for it possesses concurrently several important properties, such as being both a metric and a normalized measure, admitting an exact analytical adjusted-for-chance form, and using the nominal $[0,1]$ range better than other normalized variants.

Keywords: clustering comparison, information theory, adjustment for chance, normalized information distance

1. Introduction

Clustering comparison measures play an important role in cluster analysis. Most often, such measures are used for external validation, that is, assessing the goodness of clustering solutions according to a "ground truth" clustering. Recent advances in cluster analysis have driven new algorithms, in which the clustering comparison measures are used actively in searching for good clustering solutions. One such example occurs in the context of ensemble (consensus) clustering, whose aim is to unify a set of clusterings, already obtained by some algorithms, into a single high quality one (Singh et al., 2009; Strehl and Ghosh, 2002; Charikar et al., 2003). A possible approach is to choose the clustering which shares the most information with all the other clusterings, such as in Strehl and Ghosh (2002). A measure is therefore needed to quantify the amount of information shared between clusterings, more specifically in this case, between the "centroid" clustering and all the

∗. Also in ATP Laboratory, National ICT Australia (NICTA).
†. Also in the Victoria Research Laboratory, National ICT Australia (NICTA).

other clusterings. Another example is in model selection by stability assessment (Ben-David et al., 2006; Shamir and Tishby, 2008). A possible realization of this scheme is to measure the average pairwise distances between all the clusterings obtained under some sort of perturbations (Vinh and Epps, 2009), hence requiring a clustering comparison measure.

Numerous measures for comparing clusterings have been proposed. Besides the class of *pair-counting based* and *set-matching based* measures, *information theoretic* measures form another fundamental class. In the clustering literature, such measures have been employed because of their strong mathematical foundation, and ability to detect non-linear similarities. For the particular purpose of clustering comparison, this class of measures has been popularized through the works of Strehl and Ghosh (2002) and Meilă (2005), and since then has been employed in various subsequent research (Fern and Brodley, 2003; He et al., 2008; Asur et al., 2007; Tumer and Agogino, 2008). In this context, the pioneering works of Meilă (2003, 2005, 2007) have shown a number of desirable theoretical properties of one of these measures—the *variation of information* (VI)—such as its metric property and its alignment with the lattice of partitions. Although having received considerable interest, in our opinion, the application of information theoretic measures for comparing clustering has been somewhat scattered. Apart from the VI which possesses a fairly comprehensive characterization, less is known about the *mutual information* and various forms of the so-called *normalized mutual information* (Strehl and Ghosh, 2002). The main technical contributions of this paper can be summarized as being three-fold:

1. We first review and make a coherent categorization of information theoretic similarity and distance measures for clustering comparison. We then discuss and prove their two important properties, namely the normalization and the metric properties. We show that among the prospective measures, the *normalized information distance* (NID) and the *normalized variation of information* (NVI) satisfy both these desirable properties.

2. We draw the attention of the clustering community towards the necessity of correcting information theoretic measures for chance in certain situations, derive analytical forms for the proposed adjusted-for-chance measures, and investigate their properties. Preliminary results regarding correcting information theoretic measures for chance have previously appeared in Vinh, Epps, and Bailey (2009). In this paper, we further analyze the large sample properties of the adjusted measures, and give a recommendation as to when adjustment is mostly needed.

3. Of the available information theoretic measures, we advocate the normalized information distance (NID) as a general purpose measure for comparing clusterings, which has the advantage of being both a metric and a normalized measure, admitting an exact analytical adjusted-for-chance form, and using better the nominal $[0,1]$ range. For ease of reading, we present the proofs of all results herein in the Appendix.

2. A Brief Review of Measures for Comparing Clusterings

Let S be a set of N data items, then a (partitional) clustering $\mathbf{U}$ on S is a way of partitioning S into non-overlap subsets $\{U_1, U_2, \ldots, U_R\}$, where $\cup_{i=1}^{R} U_i = S$ and $U_i \cap U_j = \emptyset$ for $i \neq j$. The information on the overlap between two clusterings $\mathbf{U} = \{U_1, U_2, \ldots, U_R\}$ and $\mathbf{V} = \{V_1, V_2, \ldots, V_C\}$ can be summarized in form of a $R \times C$ *contingency table* $M = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}$ as illustrated in Table 1, where n_{ij} denotes the number of objects that are common to clusters U_i and V_j.

Pair counting based measures are built upon counting pairs of items on which two clusterings agree or disagree. Specifically, the $\binom{N}{2}$ item pairs in S can be classified into one of the 4 types—N_{11}:

$\mathbf{U}\backslash\mathbf{V}$	V_1	V_2	$\dots$	V_C	Sums
U_1	n_{11}	n_{12}	$\dots$	n_{1C}	a_1
U_2	n_{21}	n_{22}	$\dots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	$\dots$	n_{RC}	a_R
Sums	b_1	b_2	$\dots$	b_C	$\sum_{ij} n_{ij} = N$

Table 1: The Contingency Table, $n_{ij} = |U_i \cap V_j|$

the number of pairs that are in the same cluster in both **U** and **V**; N_{00}: the number of pairs that are in different clusters in both **U** and **V**; N_{01}: the number of pairs that are in the same cluster in **U** but in different clusters in **V**; and N_{10}: the number of pairs that are in different clusters in **U** but in the same cluster in **V**—that can be calculated using the n_{ij}'s (Hubert and Arabie, 1985). Intuitively, N_{11} and N_{00} can be used as indicators of agreement between **U** and **V**, while N_{01} and N_{10} can be used as disagreement indicators. A well known index of this class is the Rand index (RI, Rand 1971), defined straightforwardly as $\mathrm{RI}(\mathbf{U},\mathbf{V}) = (N_{00}+N_{11})/\binom{N}{2}$, which lies in the nominal range of [0,1]. In practice however, the RI often lies within the narrower range of [0.5,1]. Also, its baseline value can be high and does not take on a constant value. For these reasons, the RI has been mostly used in it adjusted form, known as the adjusted Rand index (ARI, Hubert and Arabie 1985):

$$\mathrm{ARI}(\mathbf{U},\mathbf{V}) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00}+N_{01})(N_{01}+N_{11}) + (N_{00}+N_{10})(N_{10}+N_{11})}.$$

The ARI is bounded above by 1, and equals 0 when the RI equals its expected value (under the generalized hypergeometric distribution assumption for randomness). Besides the ARI, there are many other, possibly less popular, measures in this class. Albatineh et al. (2006) made a comprehensive list of 22 different indices of this type, a number which is large enough to make the task of choosing an appropriate measure difficult and confusing. Their work, and subsequent extension of Warrens (2008), showed that after correction for chance, some of these measures become equivalent. Despite the existence of numerous measures, the ARI remains the most well-known and widely used (Steinley, 2004). Therefore, in this work, we take it as the representative of this class for comparison with other measures. Although the ARI has been mainly used in its similarity form, it can be easily shown that its distance version, that is, $1 - \mathrm{ARI}$, is not a proper metric.

Set matching based measures, as their name suggests, are based on finding matches between clusters in the two clusterings. A popular measure is the classification error rate which is often employed in supervised learning. Several other indices are discussed in Meilă (2007), all suffering from two major problems which have long been known in the clustering comparing literature (Dom, 2001; Steinley, 2004; Meilă, 2007) namely: (i) the number of clusters in the two clusterings may be different, making this approach problematic, since there are some clusters which are put outside consideration; and (ii) even when the numbers of clusters are the same, the unmatched part of each matched cluster pair is still put outside consideration. Due to the problems with this class of indices, we shall not consider them further in this paper.

Information theoretic based measures are built upon fundamental concepts from information theory (Cover and Thomas, 1991). Given two clusterings **U** and **V**, their entropies, joint entropy,

conditional entropies and mutual information (MI) are defined naturally via the marginal and joint distributions of data items in $\mathbf{U}$ and $\mathbf{V}$ respectively as:

$$\begin{aligned} H(\mathbf{U}) &= -\sum_{i=1}^{R} \frac{a_i}{N} \log \frac{a_i}{N}, \\ H(\mathbf{U},\mathbf{V}) &= -\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}}{N}, \\ H(\mathbf{U}|\mathbf{V}) &= -\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{b_j/N}, \\ I(\mathbf{U},\mathbf{V}) &= \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{a_i b_j/N^2}. \end{aligned}$$

The MI measures the information that $\mathbf{U}$ and $\mathbf{V}$ share: it tells us how much knowing one of these clusterings reduces our uncertainty about the other. From a communication theory point of view, the above-defined quantities can be interpreted as follows. Suppose we need to transmit all the cluster labels in $\mathbf{U}$ on a communication channel, then $H(\mathbf{U})$ can be interpreted as the average amount of information, for example, in bits, needed to encode the cluster label of each data point according to $\mathbf{U}$. Now suppose that $\mathbf{V}$ is made available to the receiver, then $H(\mathbf{U}|\mathbf{V})$ denotes the average number of bits needed to transmit each label in $\mathbf{U}$ if $\mathbf{V}$ is already known. We are interested in how seeing how much $H(\mathbf{U}|\mathbf{V})$ is smaller than $H(\mathbf{U})$, that is, how much the knowledge of $\mathbf{V}$ helps us to reduce the number of bits needed to encode $\mathbf{U}$. This can be quantified in terms of the mutual information $H(\mathbf{U}) - H(\mathbf{U}|\mathbf{V}) = I(\mathbf{U},\mathbf{V})$. The knowledge of $\mathbf{V}$ thus helps us to reduce the number of bits needed to encode each cluster label in $\mathbf{U}$ by an amount of $I(\mathbf{U},\mathbf{V})$ bits. In the reverse direction we also have $I(\mathbf{U},\mathbf{V}) = H(\mathbf{V}) - H(\mathbf{V}|\mathbf{U})$. Clearly, the higher the MI, the more useful the information in $\mathbf{V}$ helps us to predict the cluster labels in $\mathbf{U}$ and vice-versa.

Before closing this section, we list several generally desirable properties of a clustering comparison measure. This list is not meant to be exhaustive, and particular applications might require other specific properties.

- *Metric property*: the metric property requires that a distance measure satisfy the properties of a true metric, namely positive definiteness, symmetry and triangle inequality. As the most basic benefit, the metric property conforms to our intuition of distance (Meilă, 2007). Furthermore, it is important if one would like to study, either the structure of, or design algorithms for the complex space of clusterings, as many nice theoretical results already exist for metric spaces.

- *Normalization*: the normalization property requires that the range of a similarity or distance measure lies within a fixed range, for example, [-1,1] or [0,1]. Normalization facilitates interpretation and comparison across different conditions (Strehl and Ghosh, 2002; Luo et al., 2009), where unbounded measures might have different ranges. Also, normalization has been shown to improve the sensitiveness of certain measures, such as the MI, with respect to the difference in cluster distribution in the two clusterings (Wu et al., 2009). The fact that all of the 22 different pair counting based measures discussed in Albatineh et al. (2006) are normalized, further stresses the particular interest of the clustering community in this property.

- *Constant baseline property*: for a similarity measure, its expected value between pairs of independent clusterings, for example, clusterings sampled independently at random, should be a constant. Ideally this baseline value should be zero, indicating no similarity. The Rand index is an example of a similarity index which does not satisfy this rather intuitive property, the reason why it has been mainly used in its adjusted form.

3. Information Theoretic Based Measures - Variants and Properties

Name	Expression	Range	Related sources
Mutual Information (MI)	$I(\mathbf{U},\mathbf{V})$	$[0,\min\{H(\mathbf{U}),H(\mathbf{V})\}]$	Banerjee et al. (2005)
Normalized MI (NMI)			
NMI_{joint}	$\frac{I(\mathbf{U},\mathbf{V})}{H(\mathbf{U},\mathbf{V})}$	[0,1]	Yao (2003)
NMI_{max}	$\frac{I(\mathbf{U},\mathbf{V})}{\max\{H(\mathbf{U}),H(\mathbf{V})\}}$	[0,1]	Kvalseth (1987)
NMI_{sum}	$\frac{2I(\mathbf{U},\mathbf{V})}{H(\mathbf{U})+H(\mathbf{V})}$	[0,1]	Kvalseth (1987)
NMI_{sqrt}	$\frac{I(\mathbf{U},\mathbf{V})}{\sqrt{H(\mathbf{U})H(\mathbf{V})}}$	[0,1]	Strehl and Ghosh (2002)
NMI_{min}	$\frac{I(\mathbf{U},\mathbf{V})}{\min\{H(\mathbf{U}),H(\mathbf{V})\}}$	[0,1]	Kvalseth (1987) Liu et al. (2008)
Adjusted-for-Chance MI (see Section 4)			
$\mathrm{AMI}_{max}{}^{\dagger}$	$\frac{I(\mathbf{U},\mathbf{V})-E\{I(\mathbf{U},\mathbf{V})\}}{\max\{H(\mathbf{U}),H(\mathbf{V})\}-E\{I(\mathbf{U},\mathbf{V})\}}$	$[0,1]^{*}$	
$\mathrm{AMI}_{sum}{}^{\dagger}$	$\frac{I(\mathbf{U},\mathbf{V})-E\{I(\mathbf{U},\mathbf{V})\}}{\frac{1}{2}[H(\mathbf{U})+H(\mathbf{V})]-E\{I(\mathbf{U},\mathbf{V})\}}$	$[0,1]^{*}$	
$\mathrm{AMI}_{sqrt}{}^{\dagger}$	$\frac{I(\mathbf{U},\mathbf{V})-E\{I(\mathbf{U},\mathbf{V})\}}{\sqrt{H(\mathbf{U})H(\mathbf{V})}-E\{I(\mathbf{U},\mathbf{V})\}}$	$[0,1]^{*}$	
$\mathrm{AMI}_{min}{}^{\dagger}$	$\frac{I(\mathbf{U},\mathbf{V})-E\{I(\mathbf{U},\mathbf{V})\}}{\min\{H(\mathbf{U}),H(\mathbf{V})\}-E\{I(\mathbf{U},\mathbf{V})\}}$	$[0,1]^{*}$	

*These measures are normalized in a stochastic sense, being equal to 1 if the (unadjusted) measures equal their value as expected by chance agreement. †Our proposed measures.

Table 2: Information theoretic-based similarity measures

Similarity measures: the mutual information (MI), a non-negative quantity, can be employed as the most basic similarity measure. Based on the observation that the MI is upper-bounded by the following quantities:

$$I(\mathbf{U},\mathbf{V}) \leq \min\{H(\mathbf{U}),H(\mathbf{V})\} \leq \sqrt{H(\mathbf{U})H(\mathbf{V})} \leq \frac{1}{2}(H(\mathbf{U})+H(\mathbf{V})) \leq \max\{H(\mathbf{U}),H(\mathbf{V})\} \leq H(\mathbf{U},\mathbf{V}), \quad (1)$$

we can derive several normalized versions of the mutual information (NMI) as listed in Table 2. All the normalized variants are bounded in [0,1], equaling 1 when the two clusterings are identical, and 0 when they are independent, that is, sharing no information about each other. In the latter case, the contingency table takes the form of the so-called "independence table" where $n_{ij} = |U_i||V_j|/N$ for all i, j. The MI and some of its normalized versions have been used in the clustering literature as similarity measures between objects in general (see, for example, Yao, 2003 and references therein). For the particular purpose of clustering comparison, Banerjee et al. (2005) employed the unnormalized MI. Strehl and Ghosh (2002) on the other hand made use of the NMI_{sqrt} normalized version, which has also been used in several follow-up works in the context of ensemble clustering (Fern and Brodley, 2003; He et al., 2008; Asur et al., 2007; Tumer and Agogino, 2008).

Name	Expression	Range	Metric	Related sources
Unnormalized distance measures				
D_{joint} (Variation of Information)	$H(\mathbf{U},\mathbf{V}) - I(\mathbf{U},\mathbf{V})$	[0,log$N$]	✓	Yao (2003) Meilă (2005)
D_{max}	$\max\{H(\mathbf{U}),H(\mathbf{V})\} - I(\mathbf{U},\mathbf{V})$	[0,log$N$]	✓	
$D_{sum}(\equiv 1/2 D_{joint})$	$\frac{1}{2}[H(\mathbf{U})+H(\mathbf{V})] - I(\mathbf{U},\mathbf{V})$	[0,log$N$]	✓	
D_{sqrt}	$\sqrt{H(\mathbf{U})H(\mathbf{V})} - I(\mathbf{U},\mathbf{V})$	[0,log$N$]	✗	
D_{min}	$\min\{H(\mathbf{U}),H(\mathbf{V})\} - I(\mathbf{U},\mathbf{V})$	[0,log$N$]	✗	
Normalized distance measures				
d_{joint} (Normalized VI)	$1-\frac{I(\mathbf{U},\mathbf{V})}{H(\mathbf{U},\mathbf{V})}$	[0,1]	✓	Kraskov et al. (2005)
d_{max} (Normalized Information Distance)	$1-\frac{I(\mathbf{U},\mathbf{V})}{\max\{H(\mathbf{U}),H(\mathbf{V})\}}$	[0,1]	✓	Kraskov et al. (2005)
d_{sum}	$1-\frac{2I(\mathbf{U},\mathbf{V})}{H(\mathbf{U})+H(\mathbf{V})}$	[0,1]	✗	
d_{sqrt}	$1-\frac{I(\mathbf{U},\mathbf{V})}{\sqrt{\{H(\mathbf{U}),H(\mathbf{V})\}}}$	[0,1]	✗	
d_{min}	$1-\frac{I(\mathbf{U},\mathbf{V})}{\min\{H(\mathbf{U}),H(\mathbf{V})\}}$	[0,1]	✗	
Adjusted-for-Chance distance measures (see Section 4)				
$Ad_{max}{}^{\dagger}$	$1-\mathrm{AMI}_{max}$	$[0,1]^*$	✗	
$Ad_{sum}{}^{\dagger}$	$1-\mathrm{AMI}_{sum}$	$[0,1]^*$	✗	
$Ad_{sqrt}{}^{\dagger}$	$1-\mathrm{AMI}_{sqrt}$	$[0,1]^*$	✗	
$Ad_{min}{}^{\dagger}$	$1-\mathrm{AMI}_{min}$	$[0,1]^*$	✗	

*These measures are normalized in a stochastic sense, being equal to 0 if the (unadjusted) measures equal their value as expected by chance agreement. †Our proposed measures. D denotes an unnormalized distance measure, d denotes a normalized distance measure

Table 3: Information theoretic-based distance measures

Distance measures: based on the five upper bounds for $I(\mathbf{U},\mathbf{V})$ given in (1), we can define five distance measures, namely $D_{joint}, D_{max}, D_{sum}, D_{sqrt}$ and D_{min}, as detailed in Table 3. However, it can be seen that $D_{joint} = 2D_{sum}$,[1] and these two measures have been known in the clustering literature as the variation of information—VI (Meilă, 2005). The fact that D_{joint} (and hence D_{sum}) is a true metric is a well known result (Meilă, 2005). In addition, we also present the following new results (see Appendix for proof):

Theorem 1 *D_{max} is a metric.*

Theorem 2 *D_{min} and D_{sqrt} are **not** metrics.*

The negative result given in Theorem 2 is indeed helpful in narrowing our search scope for a reasonable distance measure. So far, D_{max} and D_{joint} (D_{sum}) are potential candidates. These distance measures do not have a fixed upper bound however, and we are therefore seeking some normalized variants. By dividing each distance measure by its corresponding upper bound we can define five normalized variants as detailed in Table 3, which are actually the unit-complements of the corresponding NMI variants, for example, $d_{joint} = 1 - \mathrm{NMI}_{joint}$. We now state the following properties of the normalized distance measures:

Theorem 3 *The normalized variation of information, d_{joint}, is a metric*

Theorem 4 *The normalized information distance, d_{max}, is a metric*

1. $2D_{sum}(\mathbf{U},\mathbf{V}) = H(\mathbf{U})+H(\mathbf{V})-2I(\mathbf{U},\mathbf{V}) = [H(\mathbf{U})+H(\mathbf{V})-I(\mathbf{U},\mathbf{V})] - I(\mathbf{U},\mathbf{V}) = H(\mathbf{U},\mathbf{V}) - I(\mathbf{U},\mathbf{V}) = D_{joint}(\mathbf{U},\mathbf{V})$.

Theorem 5 *The normalized distance measures d_{min}, d_{sum} and d_{sqrt}, are **not** metrics.*

The proofs for Theorem 3 and 4 was presented in an unofficially extended version of Kraskov et al. (2005).[2] Unfortunately, their proof for Theorem 4 was erroneous.[3] Since these are two interesting results, we give our shortened proof for Theorem 3 and a corrected proof for Theorem 4 in the Appendix. The negative results in Theorem 5 are again useful in narrowing our scope looking for a good candidate. From our discussion so far, we can now identify two promising candidates: d_{joint} and d_{max}. Since the variation of information—D_{joint}—is the unnormalized version of d_{joint}, we shall name d_{joint} the normalized variation of information (NVI). d_{max} has not been named in the literature, therefore we name it after its well known analogue in Kolmogorov complexity theory (Li et al., 2004), the normalized information distance (NID). Both the NVI and NID have the remarkable property of being both a metric and a normalized measure. We note that Meilă (2007) proposed normalized variants for the VI, such as: $V(\mathbf{U},\mathbf{V}) = \frac{1}{\log N}\mathrm{VI}(\mathbf{U},\mathbf{V})$ or: $V_{K^*}(\mathbf{U},\mathbf{V}) = \frac{1}{2\log K^*}\mathrm{VI}(\mathbf{U},\mathbf{V})$ when the number of clusters in both $\mathbf{U}$ and $\mathbf{V}$ is bounded by the same constant $K^* < \sqrt{N}$. The bounds of $\log N$ and $2\log K^*$ are not as strict as $H(\mathbf{U},\mathbf{V})$ however,[4] thus the useful range of these normalized VI variants is narrower than that of d_{joint}. The joint entropy $H(\mathbf{U},\mathbf{V})$ provides a stricter upper bound, enabling d_{joint} to better exploit the [0,1] range, while still retaining the metric property. It is noted that since $\max\{H(\mathbf{U}),H(\mathbf{V})\}$ is yet a tighter upper bound for $\mathrm{MI}(\mathbf{U},\mathbf{V})$ than $H(\mathbf{U},\mathbf{V})$, d_{max} is generally more preferable to d_{joint} since it can even better use the nominal range of $[0,1]$. A subtle point regarding normalization by quantities such as $\max\{H(\mathbf{U}),H(\mathbf{V})\}$ and $H(\mathbf{U},\mathbf{V})$, as has been brought to our attention by the Editor, is their potential side effects on the normalization process. For validation purpose for example, if $\mathbf{U}$ is the ground-truth, and $\mathbf{V}$ is the clustering obtained by some algorithm, then the normalization also depends on $\mathbf{V}$. Thus, while random quantities such as $\max\{H(\mathbf{U}),H(\mathbf{V})\}$ and $H(\mathbf{U},\mathbf{V})$ provide tighter bounds, their effect on the normalization process is not as clear as looser, fixed bounds such as $\log N$ and $2\log K^*$.

4. Adjustment for Chance

In this section we inspect the proposed information theoretic measures with respect to the third desirable property, that is, the *constant baseline property*. We shall first point out that, just like the well-known Rand index, the baseline value of information theoretic measures does not take on a constant value, and thus adjustment for chance will be needed in certain situations. Let us consider the following two motivating examples:

1) Example 1 - Distance to a "true" clustering: given a ground-truth clustering $\mathbf{U}$ with K_{true} clusters, we need to assess the goodness of two clusterings $\mathbf{V}$ with C clusters, and $\mathbf{V}'$ with C' clusters. If $C = C'$ then the situation would be quite simple. Since the setting is the same for both $\mathbf{V}$ and $\mathbf{V}'$, we expect the comparison to be "fair" under any particular measure. However if $C \neq C'$, the situation becomes more complicated. We set up an experiment as follows: consider a set of N data points, let the number of clusters K vary from 2 to K_{max} and suppose that the true clustering has $K_{true} = [K_{max}/2]$ clusters. Now for each value of K, generate 10,000 random clusterings and calculate the average MI, NMI_{max}, VI, RI and ARI between those clusterings to a fixed, random

2. Available online at `http://arxiv.org/abs/q-bio/0311039v2`.
3. In their case 1, $D'(Z,Y)$ is in fact not equal to $H(Z|Y)/H(Y)$.
4. If $N \leq RC$ then $H(\mathbf{U},\mathbf{V}) \leq \log N$, with the equality attained only when cells of the contingency table contain only either 1 or 0. If $N > RC$ then $H(\mathbf{U},\mathbf{V}) \leq \log(RC) \leq \log(K^*K^*) = 2\log K^*$.

clustering chosen as the "true" clustering. The results for two combinations of (N, K_{true}) are given in Fig. 1(a,b). It can be observed that the unadjusted measures such as the RI, MI and NMI (VI) monotonically increase (decreases) as K increases. Thus even by selecting totally at random, a 7-cluster solution would have a greater chance to outperform a 3-cluster solution, although there isn't any difference in the clustering generation methodology. A corrected-for-chance measure, such as the ARI, on the other hand, has a baseline value always close to zero, and appears not to be biased in favor of any particular value of K. The same issue is observed with all other variants of the NMI (data not shown). Thus for this example, an adjusted-for-chance version of the MI is desirable.

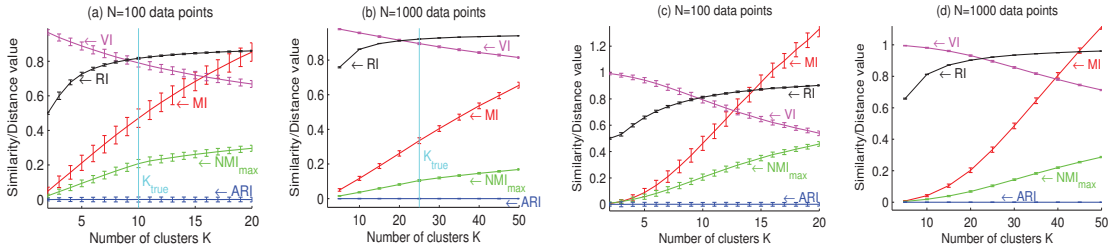

Figure 1: (a,b) Average distance between sets of random clusterings to a "true" clustering (c,d) Average pairwise distance in a set of random clusterings. Error bars denote standard deviation.

2) Example 2 - Determining the number of clusters via consensus (ensemble) clustering: in an era where a huge number of clustering algorithms exist, the consensus clustering idea (Monti et al., 2003; Strehl and Ghosh, 2002; Yu et al., 2007) has recently received increasing interest. Consensus clustering is not just another clustering algorithm: it rather provides a framework for unifying the knowledge obtained from other algorithms. Given a data set, consensus clustering employs one or several clustering algorithms to generate a set of clustering solutions on either the original data set or its perturbed versions. From these clustering solutions, consensus clustering aims to choose a robust and high quality representative clustering. Although the main objective of consensus clustering is to discover a high quality cluster structure, closer inspection of the set of clusterings obtained can often give valuable information about the appropriate number of clusters present. More specifically, we have empirically observed the following: in regard to the set of clusterings obtained, when the specified number of clusters coincides with the true number of clusters, this set has a tendency to be less diverse. This is an indication of the robustness of the obtained cluster structure. To quantify this diversity we have recently developed a novel index (Vinh and Epps, 2009), namely the *consensus index* (CI), which is built upon a suitable clustering similarity measure. Given a value of K, suppose we have generated a set of B clustering solutions $\mathcal{U}_K = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_B\}$, each with K clusters. We define the consensus index of $\mathcal{U}_K$ as:

$$\mathrm{CI}(\mathcal{U}_K) = \frac{\sum_{i<j} \mathrm{AM}(\mathbf{U}_i, \mathbf{U}_j)}{B(B-1)/2}$$

where the agreement measure AM is a suitable clustering similarity index. Thus, the CI quantifies the average pairwise agreement in $\mathcal{U}_K$. The optimal number of clusters K^* is chosen as which that maximizes CI, that is, $K^* = \arg\max_{K=2\ldots K_{max}} \mathrm{CI}(\mathcal{U}_K)$. In this setting, a normalized measure is preferable for its equalized range at different values of K. We performed an experiment as follows:

given N data points, randomly assign each data point into one of the K clusters with equal probability and check to ensure that the final clustering contains exactly K clusters. For each K, repeat this 200 times to create 200 random clusterings, then calculate the average values of the MI, VI, NMI_{max}, RI and ARI between all 19,900 clustering pairs. Typical experimental results are presented in Fig. 1(c,d). It can be observed that for a given data set, the average MI, NMI and RI (VI) values between random clusterings tend to increase (decrease) as the number of clusters increases, while the average value of the ARI is always close to zero. When the ratio of N/K is large, the average value for NMI is reasonably close to zero, but grows as N/K becomes smaller. This is clearly an unwanted effect, since a consensus index built upon the MI, NMI and VI would be biased in favour of a larger number of clusters. Thus in this situation, an adjusted-for-chance version of the MI is again important.

4.1 The Proposed Adjusted Measures

To correct the measures for randomness it is necessary to specify a model according to which random partitions are generated. Such a common model is the "permutation model" (Lancaster, 1969, p. 214), in which clusterings are generated randomly subject to having a fixed number of clusters and points in each clusters. Using this model, which was also adopted by Hubert and Arabie (1985) for the ARI, we have previously shown (Vinh et al., 2009) that the expected mutual information between two clusterings $\mathbf{U}$ and $\mathbf{V}$ is:

$$\mathbf{E}\{I(\mathbf{U},\mathbf{V})\} = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{n_{ij}=\max(a_i+b_j-N,0)}^{\min(a_i,b_j)} \frac{n_{ij}}{N}\log\left(\frac{N.n_{ij}}{a_i b_j}\right)\frac{a_i!b_j!(N-a_i)!(N-b_j)!}{N!n_{ij}!(a_i-n_{ij})!(b_j-n_{ij})!(N-a_i-b_j+n_{ij})!}. \quad (2)$$

As suggested by Hubert and Arabie (1985), the general form of a similarity index corrected for chance is given by:

$$\text{Adjusted\_Index} = \frac{\text{Index} - \text{ Expected\_Index}}{\text{Max\_Index} - \text{Expected\_Index}}, \quad (3)$$

which is upper-bounded by 1 and equals 0 when the index equals its expected value. Having calculated the expectation of the MI, we propose the adjusted form, which we call the *adjusted mutual information* (AMI), for the normalized mutual information according to (3). For example, taking the NMI_{max} we have:

$$\text{AMI}_{max}(\mathbf{U},\mathbf{V}) = \frac{\text{NMI}_{max}(\mathbf{U},\mathbf{V}) - \mathbf{E}\{\text{NMI}_{max}(\mathbf{U},\mathbf{V})\}}{1 - \mathbf{E}\{\text{NMI}_{max}(\mathbf{U},\mathbf{V})\}} = \frac{I(\mathbf{U},\mathbf{V}) - \mathbf{E}\{I(\mathbf{U},\mathbf{V})\}}{\max\{H(\mathbf{U}),H(\mathbf{V})\} - \mathbf{E}\{I(\mathbf{U},\mathbf{V})\}}.$$

Similarly, other adjusted *similarity* measures are listed in Table 2. It can be seen that the adjusted-for-chance forms of the MI are all normalized in a stochastic sense. Specifically, the AMI equals 1 when the two clusterings are identical, and 0 when the MI between the two clusterings equals its expected value. The adjusted forms for the *distance* measures, listed in Table 2, are again the unit-complements of the corresponding adjusted *similarity* measures, for example, $Ad_{max} = 1 - \text{AMI}_{max}$, and are also normalized in a stochastic sense. Following the naming scheme that we have adopted throughout in this paper, we name Ad_{max} the adjusted information distance. It is noted that at this stage, we have not been able to derive an analytical solution for the adjusted form for the normalized variation of information (d_{joint}) measure. The derivation of the expected value for this measure appears to be more involved observing that $I(\mathbf{U},\mathbf{V})$ is present in both the numerator and denominator ($H(\mathbf{U},\mathbf{V}) = H(\mathbf{U}) + H(\mathbf{V}) - I(\mathbf{U},\mathbf{V})$). We repeat the experiments described in examples 1 and 2, this time with the adjusted version of the NMI_{max}. Now it can be seen from Fig. 2 that just like the ARI, the AMI_{max} baseline values are close to zero. It is noted that in these experiments, we did not

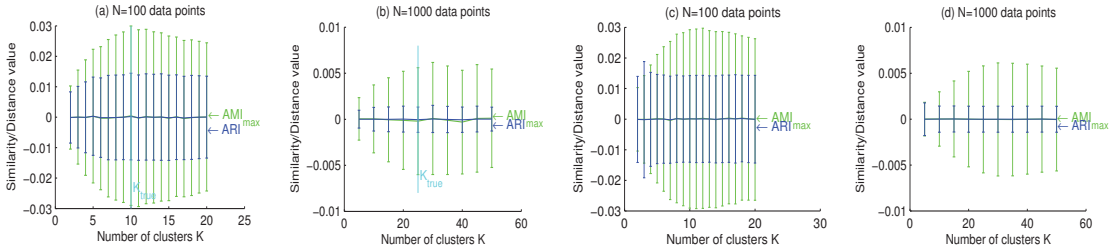

Figure 2: (a,b) Average distance between sets of random clusterings to a "true" clustering (c,d) Average pairwise distance in a set of random clusterings. Error bars denote standard deviation.

require the marginals of the contingency table to be fixed as per the assumption of the generalized hypergeometric model of randomness. Nevertheless, the adjusted measures still exhibit the desired behavior.

4.2 Properties of the Adjusted Measures

While admitting a constant baseline, the proposed adjusted-for-chance measures are, unfortunately, not proper metrics:

Theorem 6 *The adjusted measures $Ad_{max}, Ad_{sum}, Ad_{sqrt}$ and Ad_{min} are **not** metrics.*

There is thus a trade-off between the metric property and correction for chance, and the user should decide which property is of higher priority. Fortunately, during our experiments with the AMI, we have observed that when the data contain a fairly large number of items as compared to the number of clusters, for example, $N/K \geq 100$, then the expected mutual information is fairly close to zero, as can be seen in Fig. 1, suggesting scenarios where adjustment for chance is not of utmost necessity. The following results formalize this observation:

Theorem 7 *Some upper bounds for the expected mutual information between two random clusterings $\boldsymbol{U}$ and $\boldsymbol{V}$ (on a data set of N data items, with R and C clusters respectively), under the hypergeometric distribution model of randomness are given by the followings:*

$$\boldsymbol{E}\{I(\boldsymbol{U},\boldsymbol{V})\} \leq \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{a_i b_j}{N^2} \log\left(\frac{N(a_i-1)(b_j-1)}{(N-1)a_i b_j} + \frac{N}{a_i b_j}\right) \leq \log\left(\frac{N+RC-R-C}{N-1}\right). \quad (4)$$

These bounds shed light on the large sample property of the adjusted measures. The following result trivially follows:

Corollary 1 *Given R and C fixed, $\lim_{N\to\infty} \boldsymbol{E}\{I(\boldsymbol{U},\boldsymbol{V})\} = 0$, and thus the adjusted measures tend toward the normalized measures.*

Also, these bounds give useful information on whether adjustment for chance is needed. For example, on a data set of 100 data items and two clusterings $\mathbf{U}$ and $\mathbf{V}$, each having 10 clusters with sizes of $[10,10,10,10,10,10,10,10,10,10]$ and $[2,4,6,8,10,10,12,14,16,18]$ respectively, the expected MI and its upper bounds according to (2) and (4) are $\mathbf{E}\{I(\mathbf{U},\mathbf{V})\} = 0.4618 < 0.5584 <$

0.5978. As the maximum MI value is only $\log(10) = 2.3$, correction for chance is needed since the baseline is high. However, if the data size increases ten-fold to 1000 items, keeping the same number of clusters and cluster distribution, the two upper bounds are 0.0764 and 0.0780 respectively, which can be considered small enough for many applications, therefore adjustment for chance might be omitted.

4.3 An Example Application

As per our analysis, adjustment for chance for information theoretic measures is mostly needed when the number of data items is relatively small compared to the number of clusters. One such prominent example is in microarray data analysis, where biological samples are clustered using gene expression data. Due to the high cost of preparing and collecting microarray data, each class, for example, of tumor, might contain only as few as several samples. In this section we demonstrate the use of the consensus index to estimate the number of clusters in microarray data. Eight synthetic and real microarray data sets are drawn from Monti et al. (2003), as detailed in Table 4 (see the original publication for preprocessing issues). A quick check upon the (higher) upper bound of the expected MI on these data sets suggests that correction for chance will be needed, for example, on the Leukemia data set, as $K(= R = C)$ grows from 2 to 10, this upper bound grows from 0.03 to 1.16.

Simulated Data	#Classes	#Samples	#Genes	Real data	#Classes	#Samples	#Genes
Gaussian3	3	60	600	Leukemia	3	38	999
Gaussian5	5	500	2	Novartis	4	103	1000
Simulated4	4	40	600	Lung cancer	4+	197	1000
Simulated6	6-7	60	600	Normal tissues	13	90	1277

Table 4: Microarray data sets summary, source: Monti et al. (2003)

In Vinh and Epps (2009) we have shown that the CI, coupled with sub-sampling as the perturbation method, gives useful information on the appropriate number of clusters in microarray data. Herein, we experimented with random projection as the perturbation scheme. More specifically, the original data set was projected on a random set of 80% of the genes, and the K-means clustering algorithm was run with random initialization on the projected data set. For each value of K, 100 of such clustering solutions were created and the CI's for 6 measures, namely RI, ARI, MI, VI, NMI_{max} and AMI_{max} were calculated. Ideally we expect to see a strong global peak at the true number of cluster K_{true}. From Fig. 3(a) it can be observed that the unadjusted MI has a strong bias with respect to the number of clusters, increasing monotonically as K increases. Similar behavior was observed with all other data sets and therefore MI is not an appropriate measure for this purpose. For ease of presentation, we have excluded the MI from Fig. 3(b-h). The effect of adjustment for chance can be clearly observed in Fig. 3(c,d,e,h). Agreement by chance inflates the CI score of the unadjusted measures (RI, NMI, VI) in such cases, and can lead to incorrect estimation. The CI of the adjusted measures (ARI, AMI) correctly estimates the number of clusters in all synthetic data sets with high confidence, whereas on real data sets it gives correct estimations on the Leukemia and Normal tissues data set. The CI suggests only 3 clusters on the Novartis while the assumed number of clusters is 4. The Lung cancer data set is an example where human experts are not yet confident on the true number of clusters present (4+ clusters), while the CI gives a local peak at 6 clusters. These results are concordant with our previous finding (Vinh and Epps, 2009). The fact that the CI

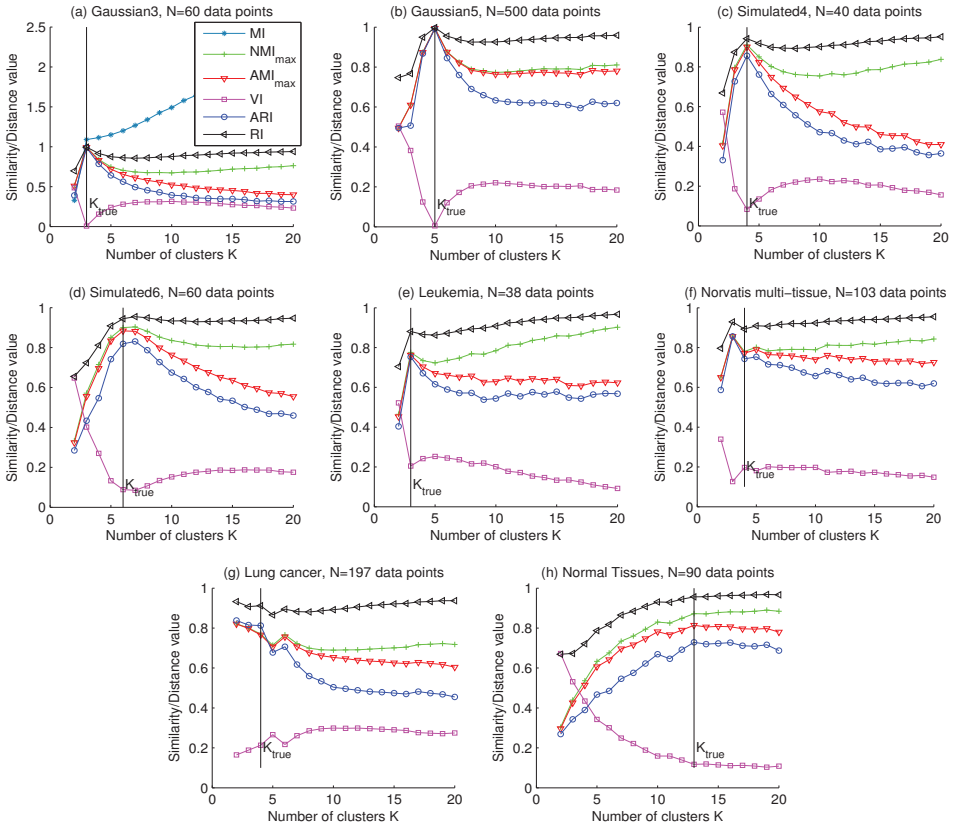

Figure 3: Consensus Index on microarray data sets.

showing global or local peaks at or near the presumably true number of clusters as attributed by the respective authors calls for further investigation on both the biological side (re-verifying the number of clusters), and the CI side (the behaviour of the index with respect to the structure of the data set, for example, the data set might contain a hierarchy of clusters and thus the CI may exhibit local peaks corresponding to such structures).

5. Related Work

Meilă (2005) considered clustering comparison measures with respect to their alignment with the lattice of partitions. In addition to the metric property, she considered three more properties namely *additivity with respect to refinement*, *additivity with respect to the join* and *convex additivity*, and showed that the VI satisfies all these properties. Unfortunately, none of the normalized or adjusted variants of the MI is fully aligned with the lattice of partitions in the above sense. Beside enhancing intuitiveness, these properties could possibly improve the efficiency of algorithms, for example, search algorithms, in the space of clusterings, though there seems not to be yet an experimental study to support such claim, calling for interesting further investigation. Nevertheless, we note that for a particular application, not always every desired property is concurrently needed at once. For example, when performing search in the space of clusterings, normalization might not be necessary, and the VI, which aligns better with the lattice of partitions, might be a more appropriate choice.

Wu et al. (2009) considered clustering comparison measures with respect to their sensitivity with class distribution. They showed that real life data can possess highly skewed class distributions, whereas some algorithms, such as K-means, tend to create balanced clusters. A good measure should therefore be sensitive to the difference in class distribution. To demonstrate this property, they used the example in Table 5, with a ground-truth clustering $\mathbf{U}$ having class sizes of [30, 2, 6, 10, 2], and two clustering solutions: $\mathbf{V}$ having cluster sizes of [10, 10, 10, 10, 10]; and $\mathbf{V}'$ having cluster sizes of [29, 2, 6, 11, 2]. It is easily seen that $\mathbf{V}'$ closely reflects the class structure in $\mathbf{U}$, and thus should be judged closer to $\mathbf{U}$ than $\mathbf{V}$. The authors showed that the unnormalized MI is a "defective measure", in that it judges $\mathrm{MI}(\mathbf{U},\mathbf{V}) > \mathrm{MI}(\mathbf{U},\mathbf{V}')$, and suggested using the "normalized VI" (d_{sum}). It can be shown that among the normalized and adjusted variants of the MI considered in this paper, only the $\mathrm{NMI}_{min}, D_{min}, d_{min}$ and Ad_{min} are defective measures in the above sense.

I	U_1	U_2	U_3	U_4	U_5	II	U_1	U_2	U_3	U_4	U_5
V_1	10	0	0	0	0	V'_1	27	0	0	2	0
V_2	10	0	0	0	0	V'_2	0	2	0	0	0
V_3	10	0	0	0	0	V'_3	0	0	6	0	0
V_4	0	0	0	10	0	V'_4	3	0	0	8	0
V_5	0	2	6	0	2	V'_5	0	0	0	0	2

Table 5: Two clustering results

6. Conclusion

This paper has presented an organized study of information theoretic measures for clustering comparison. We have shown that the normalized information distance (NID) and normalized variation of information (NVI) satisfy both the normalization and the metric properties. Between the two, the NID is preferable since the tighter upper bound of the MI used for normalization allows it to better use the [0,1] range. We highlighted the importance of correcting these measures for chance agreement, especially when the number of data points is relatively small compared with the number of clusters, for example, in the case of microarray data analysis. One of the theoretical advantages of the NID over the popular adjusted Rand index is that it can be used in the non-adjusted form (when N/K is large), thus enjoying the property of being a true metric in the space of clusterings. We therefore advocate the NID as a "general purpose" measure for clustering validation, comparison and algorithm design, for it possesses concurrently several useful and important properties. Nevertheless, we note that for a particular application scenario, not always every desired property is needed concurrently, and therefore the user should prioritize the most important property. Our research systematically organizes and complements the current literature to help readers make more informed decisions.

Acknowledgments

We thank the Action Editor and the anonymous reviewers for their constructive comments. This work was partially supported by NICTA and the Australian Research Council.

Availability: Matlab code for computing the adjusted mutual information (AMI) is available from `http://ee.unsw.edu.au/~nguyenv/Software.htm`.

Appendix A. Proofs

Proof (Theorem 1) We only prove the triangle inequality as other parts are trivial. We first show that

$$H(X|Y) \leq H(X|Z) + H(Z|Y) \tag{5}$$

holds true, since $H(X|Y) \leq H(X,Z|Y) = H(X|Z,Y) + H(Z|Y) \leq H(X|Z) + H(Z|Y)$ (the last inequality holds since conditioning always decreases entropy). We now prove the main theorem. Without loss of generality, assume that $H(Y) \leq H(X)$, and therefore $H(X|Y) \geq H(Y|X)$. The proof uses (5):

- Case 1: $H(Z) \leq H(Y)$

  $$D_{max}(X,Y) = H(X|Y) \leq H(X|Z) + H(Z|Y) \leq H(X|Z) + H(Y|Z) = D_{max}(X,Z) + D_{max}(Y,Z)$$

- Case 2: $H(Y) < H(Z) \leq H(X)$

  $$D_{max}(X,Y) = H(X|Y) \leq H(X|Z) + H(Z|Y) = D_{max}(X,Z) + D_{max}(Y,Z)$$

- Case 3: $H(X) < H(Z)$

  $$D_{max}(X,Y) = H(X|Y) \leq H(X|Z) + H(Z|Y) \leq H(Z|X) + H(Z|Y) = D_{max}(X,Z) + D_{max}(Y,Z)$$

■

Proof (Theorem 3) We prove the triangle inequality. Without loss of generality, assume that $H(X) \geq H(Y)$, therefore $H(X|Y) \geq H(Y|X)$ and $NID(X,Y) = H(X|Y)/H(X)$. The proof uses inequality (5) and simple algebra manipulations:

- Case 1: $H(Z) \leq H(Y) \leq H(X)$

  $$NID(X,Y) = \frac{H(X|Y)}{H(X)} \leq \frac{H(X|Z) + H(Z|Y)}{H(X)} \leq \frac{H(X|Z) + H(Y|Z)}{H(X)} \leq \frac{H(X|Z)}{H(X)} + \frac{H(Y|Z)}{H(Y)}$$

- Case 2: $H(Y) \leq H(Z) \leq H(X)$

  $$NID(X,Y) = \frac{H(X|Y)}{H(X)} \leq \frac{H(X|Z)}{H(X)} + \frac{H(Z|Y)}{H(X)} \leq \frac{H(X|Z)}{H(X)} + \frac{H(Z|Y)}{H(Z)} = NID(X,Z) + NID(Z,Y)$$

- Case 3: $H(Z) > H(X)$

  $$NID(X,Y) = \frac{H(X|Y)}{H(X)} < \frac{H(X|Z) + H(Z|Y)}{H(X)}$$

  If the $r.h.s \leq 1$ then adding $H(Z) - H(X) > 0$ to both its nominator and denominator will increase it:

  $$r.h.s \leq \frac{H(X|Z) + H(Z|Y) + H(Z) - H(X)}{H(X) + H(Z) - H(X)} = \frac{H(Z|X)}{H(Z)} + \frac{H(Z|Y)}{H(Z)} = NID(X,Z) + NID(Z,Y),$$

therefore the triangle inequality holds. Otherwise if the $r.h.s > 1$ then adding $H(Z)-H(X) > 0$ to both its nominator and denominator as above will decrease it, but it will still be larger than 1. Therefore we also have:

$$NID(X,Y) \leq 1 < \frac{H(X|Z)+H(Z|Y)+H(Z)-H(X)}{H(X)+H(Z)-H(X)} = NID(X,Z)+NID(Z,Y).$$

■

Proof (Theorem 4) Again only the triangle inequality proof is of interest. It is sufficient to prove the following inequality:

$$\frac{H(X|Y)}{H(X,Y)} \leq \frac{H(X|Z)}{H(X,Z)} + \frac{H(Z|Y)}{H(Z,Y)},$$

then swap X and Y to obtain another analogous inequality and add them together we have the triangle inequality. The proof uses inequality (5) and simple algebra manipulations:

$$\frac{H(X|Y)}{H(X,Y)} = \frac{H(X|Y)}{H(Y)+H(X|Y)} \leq \frac{H(X|Z)+H(Z|Y)}{H(Y)+H(X|Z)+H(Z|Y)} = \frac{H(X|Z)+H(Z|Y)}{H(X|Z)+H(Y,Z)} = \dots$$
$$= \frac{H(X|Z)}{H(X|Z)+H(Y,Z)} + \frac{H(Z|Y)}{H(X|Z)+H(Y,Z)} \leq \frac{H(X|Z)}{H(X|Z)+H(Z)} + \frac{H(Z|Y)}{H(Y,Z)} = \frac{H(X|Z)}{H(X,Z)} + \frac{H(Z|Y)}{H(Z,Y)}.$$

■

Proof (Theorems 2 and 5) It is sufficient to point out a single counter example where the triangle inequality is violated. Let X and Y be two *independent* random binary variables with probability $P(X=1)=P(X=0)=P(Y=1)=P(Y=0)=1/2$, then $Z=[X;Y]$ is also a random variable with four discrete values. It is straightforward to verify that the triangle inequality is violated for all the mentioned distance measures, for example, $D_{min}(X,Y)=1<D_{min}(X,Z)+D_{min}(Y,Z)=0$. ■

Proof (Theorem 6) For $N=5$, a counter example for the triangle inequality is the following three clusterings: $\mathbf{U}=\{U_3U_1U_1U_1U_2\}, \mathbf{V}=\{V_2V_2V_3V_1V_2\}, \mathbf{X}=\{X_2X_1X_1X_1X_2\}$.

Similarly, for $N=5+d$ $(d\in\mathbb{N}^+)$, a counter example for the triangle inequality is the following three clusterings: $\mathbf{U}=\{U_3U_1U_1U_1U_2U_6U_7\dots U_{5+d}\}$, $\mathbf{V}=\{V_2V_2V_3V_1V_2V_6V_7\dots V_{5+d}\}$, $\mathbf{X}=\{X_2X_1X_1X_1X_2X_6X_7\dots X_{5+d}\}$. ■

Proof (Theorem 7) The following facts from the generalized hypergeometric distribution will be useful:

$$\mathbf{E}(n_{ij}) = \sum_{n_{ij}} n_{ij}\mathcal{P}(M|n_{ij},a,b) = \frac{a_i,b_j}{N}, \tag{6}$$
$$\mathbf{E}(n_{ij}^2) = \sum_{n_{ij}} n_{ij}^2\mathcal{P}(M|n_{ij},a,b) = \frac{a_i(a_i-1)b_j(b_j-1)}{N(N-1)} + \frac{a_ib_j}{N},$$

where $\mathcal{P}\{M|n_{ij},a,b\} = \frac{\binom{N}{n_{ij}}\binom{N-n_{ij}}{a_i-n_{ij}}\binom{N-a_i}{b_j-n_{ij}}}{\binom{N}{a_i}\binom{N}{b_j}}$ is the probability of encountering a contingency table M having fixed marginals a,b and the $(i,j)-th$ entry being n_{ij} under the generalized hypergeometric

model of randomness. Note that for the sake of notational simplicity we have dropped the lower and upper values of n_{ij} which runs from $\max((a_i+b_j-N),0)$ to $\min(a_i,b_j)$ in the sums. From (6) we have:

$$\mathbf{E}(n_{ij}) = \sum_{n_{ij}} n_{ij}\mathcal{P}(M|n_{ij},a,b) = \frac{a_i b_j}{N}\sum_{n_{ij}} \frac{n_{ij}N}{a_i b_j}\mathcal{P}(M|n_{ij},a,b) = \frac{a_i b_j}{N},$$

therefore: $\sum_{n_{ij}} \frac{n_{ij}N}{a_i b_j}\mathcal{P}(M|n_{ij},a,b) = 1$. Let $Q(n_{ij}) = \frac{n_{ij}N}{a_i b_j}\mathcal{P}(M|n_{ij},a,b)$, then we can think of $Q(n_{ij})$ as a discrete probability distribution on n_{ij}. Applying Jensen's inequality ($\mathbf{E}(f(x)) \leq f(\mathbf{E}(x))$ for f concave) to the concave logarithm function yields:

$$\sum_{n_{ij}} \frac{n_{ij}}{N}\log(\frac{N.n_{ij}}{a_i b_j})\mathcal{P}(M|n_{ij},a,b) = \sum_{n_{ij}} \frac{a_i b_j}{N^2}\log(\frac{N.n_{ij}}{a_i b_j})Q(n_{ij}) \leq \frac{a_i b_j}{N^2}\log\left(\mathbf{E}_Q(\frac{N.n_{ij}}{a_i b_j})\right). \quad (7)$$

Now, let us calculate $\mathbf{E}_Q(\frac{N.n_{ij}}{a_i b_j})$:

$$\begin{aligned}\mathbf{E}_Q(\frac{N.n_{ij}}{a_i b_j}) &= \sum_{n_{ij}} \frac{N.n_{ij}}{a_i b_j}Q(n_{ij}) = \sum_{n_{ij}} \frac{N.n_{ij}}{a_i b_j}\frac{n_{ij}N}{a_i b_j}\mathcal{P}(M|n_{ij},a,b) = \frac{N^2}{a_i^2 b_j^2}\sum_{n_{ij}} n_{ij}^2\mathcal{P}(M|n_{ij},a,b) \\ &= \frac{N^2}{a_i^2 b_j^2}\left(\frac{a_i(a_i-1)b_j(b_j-1)}{N(N-1)} + \frac{a_i b_j}{N}\right) = \frac{N(a_i-1)(b_j-1)}{(N-1)a_i b_j} + \frac{N}{a_i b_j}.\end{aligned}$$

Substituting this expression into (7) yields:

$$\sum_{n_{ij}} \frac{n_{ij}}{N}\log(\frac{N.n_{ij}}{a_i b_j})\mathcal{P}(M|n_{ij},a,b) \leq \frac{a_i b_j}{N^2}\log\left(\frac{N(a_i-1)(b_j-1)}{(N-1)a_i b_j} + \frac{N}{a_i b_j}\right).$$

Finally:

$$\mathbf{E}\{I(\mathbf{U},\mathbf{V})\} = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{n_{ij}} \frac{n_{ij}}{N}\log(\frac{N.n_{ij}}{a_i b_j})P(M|n_{ij},a,b) \leq \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{a_i b_j}{N^2}\log\left(\frac{N(a_i-1)(b_j-1)}{(N-1)a_i b_j} + \frac{N}{a_i b_j}\right). \quad (8)$$

Note that $\sum_{ij} a_i b_j/N^2 = 1$, continue applying Jensen's inequality on (8) yields:

$$\mathbf{E}\{I(\mathbf{U},\mathbf{V})\} \leq \log\left(\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{a_i b_j}{N^2}(\frac{N(a_i-1)(b_j-1)}{(N-1)a_i b_j} + \frac{N}{a_i b_j})\right) = \log\left(\frac{N+RC-R-C}{N-1}\right)$$

■

References

A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko. On similarity indices and correction for chance agreement. *Journal of Classification*, 23(2):301–313, 2006.

S. Asur, D. Ucar, and S. Parthasarathy. An ensemble framework for clustering protein-protein interaction networks. *Bioinformatics*, 23(13):i29–i40, 2007.

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382, 2005.

S. Ben-David, U. von Luxburg, and D. Pal. A sober look at clustering stability. In *19th Annual Conference on Learning Theory (COLT 2006)*, pages 5–19, 2006.

M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. In *FOCS '03: Procs. IEEE Symposium on Foundations of Computer Science*, 2003.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

B. E. Dom. An information-theoretic external cluster-validity measure. Technical report, Research Report RJ 10219, IBM, 2001.

X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Procs. ICML'03*, pages 186–193, 2003.

Z. He, X. Xu, and S. Deng. k-anmi: A mutual information based clustering algorithm for categorical data. *Inf. Fusion*, 9(2):223–233, 2008.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classif.*, 2(1):193–218, 1985.

A. Kraskov, H. Stogbauer, R. G. Andrzejak, and P. Grassberger. Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, 70(2):278–284, 2005.

T. O. Kvalseth. Entropy and correlation: Some comments. *Systems, Man and Cybernetics, IEEE Transactions on*, 17(3):517–519, 1987.

H.O Lancaster. The chi-squared distribution. New York, 1969. John Wiley.

M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi. The similarity metric. *Information Theory, IEEE Transactions on*, 50(12):3250–3264, 2004.

Z. Liu, Z. Guo, and M. Tan. Constructing tumor progression pathways and biomarker discovery with fuzzy kernel kmeans and dna methylation data. *Cancer Inform*, 6:1–7, 2008.

P. Luo, H. Xiong, G. Zhan, J. Wu, and Z. Shi. Information-theoretic distance measures for clustering validation: Generalization and normalization. *IEEE Trans. on Knowl. and Data Eng.*, 21(9): 1249–1262, 2009.

M. Meilă. Comparing clusterings by the variation of information. In *COLT '03*, pages 173–187, 2003.

M. Meilă. Comparing clusterings: an axiomatic view. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 577–584, 2005. ISBN 1-59593-180-5.

M. Meilă. Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98(5):873–895, 2007.

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

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

O. Shamir and N. Tishby. Model selection and stability in k-means clustering. In *21th Annual Conference on Learning Theory (COLT 2008)*, 2008.

V. Singh, L. Mukherjee, J. Peng, and J. Xu. Ensemble clustering using semidefinite programming with applications. *Mach. Learn.*, 2009. doi: 10.1007/s10994-009-5158-y.

D. Steinley. Properties of the Hubert-Arabie adjusted Rand index. *Psychol Methods*, 9(3):386–96, 2004.

A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

K. Tumer and A.K. Agogino. Ensemble clustering with voting active clusters. *Pattern Recognition Letters*, 29(14):1947–1953, 2008.

N. X. Vinh and J. Epps. A novel approach for automatic number of clusters detection in microarray data based on consensus clustering. In *BIBE'09: Procs. IEEE Int. Conf. on BioInformatics and BioEngineering*, 2009.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *ICML '09*, 2009.

M. Warrens. On similarity coefficients for 2x2 tables and correction for chance. *Psychometrika*, 73 (3):487–502, 2008.

J. Wu, H. Xiong, and J. Chen. Adapting the right measures for k-means clustering. In *KDD '09*, 2009.

Y. Y. Yao. Information-theoretic measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pages 115–136. Karmeshu (ed.), Springer, 2003.

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