# Comparing Clusterings – an information based distance

Marina Meilă *

**Abstract**

This paper proposes an information theoretic criterion for comparing two partitions, or *clusterings*, of the same data set. The criterion, called variation of information (VI), measures the amount of information lost and gained in changing from clustering $\mathcal{C}$ to clustering $\mathcal{C}'$. The basic properties of VI are presented and discussed. We focus on two kinds of properties: (1) those that help one build intuition about the new criterion (in particular, it is shown the VI is a true metric on the space of clusterings), and (2) those that pertain to the comparability of VI values over different experimental conditions. As the latter properties have rarely been discussed explicitly before, other existing comparison criteria are also examined in their light. Finally we present the VI from an axiomatic point of view, showing that it is the only "sensible" criterion for comparing partitions that is both aligned to the lattice and convexely additive. As a consequence, we prove an impossibility result for comparing partitions: there is no criterion for comparing partitions that simultaneoulsly satisfies the above two desirable properties and is bounded.

*Key words:* Agreement measures; Clustering; Comparing partitions; Information theory; Mutual information; Similarity measures

* Department of Statistics, University of Washington, Box 354322, Seattle WA

## 1 Introduction

Clustering, or finding partitions [1] in data, has become an increasingly popular part of data analysis. Each day new data are clustered, new clustering criteria are introduced and new algorithms are published. Clustering as a field and its connection to statistics can be traced some decades back [1,5,7,9,10,15]. However, judged by its maturity, clustering is a young domain of research, where rigorous methodology is still striving to emerge.

In particular, to empirically assess the performance of a clustering algorithm by comparing its output to a given "correct" clustering, one needs to define a "distance" on the space of partitions of a data set. This paper introduces the *variation of information (VI)*, a function that measures the distance between two partitions of the same data set. In accordance to existing terminology, such a function is called a *criterion for comparing partitions*. The VI is not the first criterion for comparing clusterings, and section 2 presents several previously used criteria discussing some of their shortcomings, which motivated the present approach.

A related reason to consider VI, that should be kept in mind throughout the reading of the paper, is that distances like it are rarely used alone. For instance, one may use the VI or another "distance" $d$ to compare clustering algorithms. A likely scenario is that one has a data set $D$, with a given "correct" clustering. Algorithm $\mathcal{A}$ is used to cluster $D$, and the resulting clustering is compared to

98195-4322, phone: (206)543-8484, e-mail: mmp@stat.washington.edu

[1] We will use the terms *partition* and *clustering* of a set interchangeably. We also use the terms *distance* and *criterion for comparing* partitions almost interchangeably; the subtle difference between the two will appear later in the paper.

the correct one via $d$. If the algorithm $\mathcal{A}$ is not completely deterministic (e.g the result may depend on initial conditions), the operation may be repeated several times, and the resulting distances to the correct clustering may be averaged to yield the algorithm's average performance. Moreover, this average may be compared to another average distance obtained in the same way for another algorithm $\mathcal{A}'$. Thus, in practice, distances between clusterings are subject to addition, subtraction and even more complex operations as will be seen shortly. This is why we want to have a clustering comparison criterion that will license such operations, inasmuch as it makes sense in the context of the application.

Here is another scenario involving complex operations with distances. Many clustering algorithms produce partitions that depend on the algorithm's initialization. For such algorithms it is customary to repeat the procedure many times, with different initial conditions. Thus, the end result is a set of clusterings instead of a single one. Passing the whole set to the human expert for examination would defeat the purpose of data exploration. It is helpful in this context to select a subset of "most diverse" clusterings; or, alternatively, to further "cluster" the similar clusterings into groups and to then choose a representative clustering or a "consensus" clustering from each group. For such operations, a distance between clusterings is required.

We define the new distance in section 3. The VI is based on the concepts of entropy and information. These have a long standing as vehicles for formalizing intuitive notions related to uncertainty. By approaching the relationship between two clusterings from the point of view of the information exchange – loss and gain – between them, we are exploiting once again the intuitive appeal of information theoretic concepts. This is much needed, since the space

of all possible clusterings of a set, although finite, has a structure complex enough to challenge human intuition [2]. As it will be shown, the choice is also fortunate from other points of view. In particular, the variation of information is provably a metric on the space of clusterings.

To address the ill-posedness of the search for a "best" criterion, section 4 presents a variety of properties of the variation of information and discusses their meaning from the point of view of comparing clusterings. We will check whether the properties of the new criterion are "reasonable" and "desirable" in a generic setting. The reader with a particular application in mind has in these properties a precise description of the criterion's behavior.

The properties themselves, although elementary (like "convex additivity"), have rarely, if ever, been discussed in the literature on comparing clusterings. Therefore, sections 5 and 6 examine other metrics and criteria of similarity between clusterings in light of these properties.

In section 7 we derive the variation of information from axioms and examine related criteria via alterations of the axioms. On this occasion, an interesting impossibility result is obtained. The discussion in section 8 concludes the paper.

[2] The number of partitions of $n$ elements into $K$ groups is the Stirling number of the second kind $S(n, K)$ [17].

## 2 Related work

A clustering $\mathcal{C}$ is a partition of a set of points, or *data set* $D$ into mutually disjoint subsets $C_1, C_2, \ldots C_K$ called *clusters*. Formally,

$$\mathcal{C} = \{C_1, C_2, \ldots C_K\} \text{ such that } C_k \cap C_l = \emptyset \text{ and } \bigcup_{k=1}^{K} C_k = D.$$

Let the number of data points in $D$ and in cluster $C_k$ be $n$ and $n_k$ respectively. We have, of course, that

$$n = \sum_{k=1}^{K} n_k \tag{1}$$

We also assume that $n_k > 0$; in other words, that $K$ represents the number of non-empty clusters. Let a second clustering of the same data set $D$ be $\mathcal{C}' = \{C'_1, C'_2, \ldots C'_{K'}\}$, with cluster sizes $n'_{k'}$. Note that the two clusterings may have different numbers of clusters.

Virtually all criteria for comparing clustering can be described using the so-called *confusion matrix*, or *association matrix* or *contingency table* of the pair $\mathcal{C}, \mathcal{C}'$. The contingency table is a $K \times K'$ matrix, whose $kk'$-th element is the number of points in the intersection of clusters $C_k$ of $\mathcal{C}$ and $C'_{k'}$ of $\mathcal{C}'$.

$$n_{kk'} = |C_k \cap C'_{k'}|$$

### *2.1 Comparing clusterings by counting pairs*

An important class of criteria for comparing clusterings, is based on counting the pairs of points on which two clusterings agree/disagree. A pair of points from $D$ can fall under one of four cases described below.

$N_{11}$ the number of point pairs that are in the same cluster under both $\mathcal{C}$ and $\mathcal{C}'$

$N_{00}$ number of point pairs in different clusters under both $\mathcal{C}$ and $\mathcal{C}'$

$N_{10}$ number of point pairs in the same cluster under $\mathcal{C}$ but not under $\mathcal{C}'$

$N_{01}$ number of point pairs in the same cluster under $\mathcal{C}'$ but not under $\mathcal{C}$

The four counts always satisfy

$$N_{11} + N_{00} + N_{10} + N_{01} \;=\; n(n-1)/2.$$

They can be obtained from the contingency table $[n_{kk'}]$. For example $2N_{11} = \sum_{k,k'} n_{kk'}^2 - n$. See [4] for details.

Wallace [20] proposed the two asymmetric criteria $\mathcal{W}_I, \mathcal{W}_{II}$ below.

$$\mathcal{W}_I(\mathcal{C},\mathcal{C}') \;=\; \frac{N_{11}}{\sum_k n_k(n_k-1)/2} \tag{2}$$

$$\mathcal{W}_{II}(\mathcal{C},\mathcal{C}') \;=\; \frac{N_{11}}{\sum_{k'} n'_{k'}(n'_{k'}-1)/2} \tag{3}$$

They represent the probability that a pair of points which are in the same cluster under $\mathcal{C}$ (respectively $\mathcal{C}'$) are also in the same cluster under the other clustering.

Fowlkes and Mallows [4] introduced a criterion which is symmetric, and is the geometric mean of $\mathcal{W}_I, \mathcal{W}_{II}$.

$$\mathcal{F}(\mathcal{C},\mathcal{C}') \;=\; \sqrt{\mathcal{W}_I(\mathcal{C},\mathcal{C}')\mathcal{W}_{II}(\mathcal{C},\mathcal{C}')} \tag{4}$$

It can be shown that this index represents a scalar product [2].

The Fowlkes-Mallows index $\mathcal{F}$ has a base-line that is the expected value of the criterion under a null hypothesis corresponding to "independent" cluster-

ings [4]. The index is used by subtracting the base-line and normalizing by the range, so that the expected value of the normalized index is 0 while the maximum (attained for identical clusterings) is 1. Note that some pairs of clusterings may theoretically result in negative indices under this normalization.

A similar transformation was introduced by [6] for Rand's [15]

$$\mathcal{R}(\mathcal{C},\mathcal{C}') \;=\; \frac{N_{11}+N_{00}}{n(n-1)/2} \tag{5}$$

The resulting adjusted Rand index has the expression

$$\mathcal{AR}(\mathcal{C},\mathcal{C}') = \frac{\mathcal{R}(\mathcal{C},\mathcal{C}') - E[\mathcal{R}]}{1-E[\mathcal{R}]} = \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K'}\binom{n_{kk'}}{2} - \left[\sum_{k=1}^{K}\binom{n_k}{2}\right]\left[\sum_{k'=1}^{K'}\binom{n'_{k'}}{2}\right] / \binom{n}{2}}{\left[\sum_{k=1}^{K}\binom{n_k}{2}+\sum_{k'=1}^{K'}\binom{n'_{k'}}{2}\right] /2 - \left[\sum_{k=1}^{K}\binom{n_k}{2}\right]\left[\sum_{k'=1}^{K'}\binom{n'_{k'}}{2}\right] / \binom{n}{2}}$$

The main motivation for adjusting indices like $\mathcal{R}$ and $\mathcal{F}$ is the observation that the unajdusted $\mathcal{R}$, $\mathcal{F}$ do not range over the entire $[0,1]$ interval (i.e $\min \mathcal{R} > 0$, $\min \mathcal{F} > 0$). In practice, the $\mathcal{R}$ index concentrates in a small interval near 1; this situation was well illustrated by [4].

But the use of adjusted indices is not without problems. First, some researchers [20] have expressed concerns at the plausibility of the null model. Second, the value of the baseline for $\mathcal{F}$ varies sharply between near 0.6 to near 0 for $n/K > 3$. The useful range of the criterion thus varies from approximately (0,1] to approximately (0.6 1] [4]. The base-line for the adjusted Rand index, as shown by the simulations of [4], varies even more: from 0.5 to 0.95. If the base-line of a criterion varies so much over the domain where most interesting clusterings are, it makes one wonder whether any linearity can be assumed for

the remaining range of this criterion, even after the appropriate normalization. In other words, is a value $\mathcal{F} = 0.7$ at a baseline of 0.4 the same as a value of $\mathcal{F} = 0.6$ at a baseline of 0.2? Note that both values, after normalization, yield 0.5 so theoretically they should be equal.

There are other criteria in the literature, to which the above discussion applies. For instance, the Jaccard [2] index

$$\mathcal{J}(\mathcal{C}, \mathcal{C}') \;=\; \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{7}$$

and the Mirkin [13] metric

$$\mathcal{M}(\mathcal{C}, \mathcal{C}') \;=\; \sum_k n_k^2 + \sum_{k'} {n'_{k'}}^2 - 2 \sum_k \sum_{k'} n_{kk'}^2 \tag{8}$$

The latter is obviously 0 for identical clusterings and positive otherwise. In fact, this metric corresponds to the Hamming distance between certain binary vector representations of each partition [13]. This metric can also be rewritten as

$$\mathcal{M}(\mathcal{C}, \mathcal{C}') \;=\; 2(N_{01} + N_{10}) \;=\; n(n-1)[1 - \mathcal{R}(\mathcal{C}, \mathcal{C}')] \tag{9}$$

Thus the Mirkin metric is another adjusted form of the Rand index.

### *2.2 Comparing clusterings by set matching*

A second category of criteria is based on set cardinality alone and does not make any assumption about how the clusterings may have been generated.

Meilă and Heckerman [12] computed the criterion $\mathcal{H}$: First, each cluster of $\mathcal{C}$

is given a "best match" in $\mathcal{C}'$. Then, $\mathcal{H}$ is computed as the total "unmatched" probability mass in the confusion matrix. More precisely,

$$\mathcal{H}(\mathcal{C}, \mathcal{C}') \;=\; 1 - \frac{1}{n} \max_{\pi} \sum_{k=1}^{K} n_{k,\pi(k)} \tag{10}$$

In the above, it is assumed w.l.o.g that $K \leq K'$, $\pi$ is an injective mapping of $\{1, \ldots K\}$ into $\{1, \ldots K'\}$, and the maximum is taken over all such mappings. In other words, for each $\pi$ we have a (partial) correspondence between the cluster labels in $\mathcal{C}$ and $\mathcal{C}'$; now looking at clustering as a classification task with the fixed label correspondence, we compute the *classification error* of $\mathcal{C}'$ w.r.t $\mathcal{C}$. The minimum possible classification error under all correspondences is $\mathcal{H}$. The index is symmetric and takes value 1 for identical clusterings. Further properties of this index are discussed in [11,18].

[8] use

$$\mathcal{L}(\mathcal{C}, \mathcal{C}') \;=\; \frac{1}{K} \sum_{k} \max_{k'} \frac{2n_{kk'}}{n_k + n'_{k'}} \tag{11}$$

This is an asymmetric criterion that is 1 when the clusterings are identical. The asymmetry of $\mathcal{L}$ raises some difficulties. Take for example the situation where $\mathcal{C} = \{D\}$ (a single cluster) and $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting off two clusters of size $nf$ (where $0 < f << 1$) from the initial large cluster. Then,

$$\mathcal{L}(\mathcal{C}, \mathcal{C}') \;=\; 1 - 2f$$

which is reasonable, but

$$\mathcal{L}(\mathcal{C}', \mathcal{C}) \;=\; \frac{1 + 2f}{3(1 - f)}$$

The above value converges to 1/3 when $f \to 0$ against our intuition that the

FIGURE 1 GOES HERE

Fig. 1. Clustering $\mathcal{C}'$ is obtained from $\mathcal{C}$ by moving a small fraction of the points in each cluster to the next cluster; $\mathcal{C}''$ is obtained from $\mathcal{C}$ by reassigning the same fraction of each cluster equally between all other clusters. The set matching criteria $\mathcal{D}$, $\mathcal{H}$, $\mathcal{L}$ declare $\mathcal{C}'$, $\mathcal{C}''$ equidistant from the original clustering $\mathcal{C}$.

difference should be negligible for small enough $f$. $\mathcal{H}$ which is normalized by $n$ not by the number of clusters $K$, does not suffer from the above problem, yielding the intuitively acceptable value of $1 - 2f$.

A symmetric criterion that is also a metric was introduced by van Dongen [19]

$$\mathcal{D}(\mathcal{C}, \mathcal{C}') \;=\; 2n - \sum_{k} \max_{k'} n_{kk'} - \sum_{k'} \max_{k} n_{kk'} \tag{12}$$

Hence, $\mathcal{D}$ is 0 for identical clusterings and strictly smaller than $2n$ otherwise.

All three above criteria however, suffer from the "problem of matching" that we discuss now. One way or another, $\mathcal{L}, \mathcal{H}, \mathcal{D}$ all first find a "best match" for each cluster, then add up the contributions of the matches found. In doing so, the criteria completely ignore what happens to the "unmatched" part of each cluster. To make things clear, let us look at the example depicted in figure 1. Suppose $\mathcal{C}$ is a clustering with $K$ equal clusters. The clustering $\mathcal{C}''$ is obtained from $\mathcal{C}$ by moving a fraction $f$ of the points in each $C_k$ to the cluster $C_{k+1(modK)}$. The clustering $\mathcal{C}'$ is obtained from $\mathcal{C}$ by reassigning a fraction $f$ of the points in each $C_k$ evenly between the other clusters. If $f < 0.5$ then $\mathcal{L}(\mathcal{C}, \mathcal{C}') = \mathcal{L}(\mathcal{C}, \mathcal{C}'')$, $\mathcal{H}(\mathcal{C}, \mathcal{C}') = \mathcal{H}(\mathcal{C}, \mathcal{C}'')$, $\mathcal{D}(\mathcal{C}, \mathcal{C}') = \mathcal{D}(\mathcal{C}, \mathcal{C}'')$. This contradicts the intuition that $\mathcal{C}'$ is a less disrupted version of $\mathcal{C}$ than $\mathcal{C}''$.

## 3 The variation of information

Now we introduce the variation of information (VI), the criterion we propose for comparing two clusterings. This criterion does not fall in any of the two categories discussed above. Therefore, we present VI from two perspectives: in this section, we motivate the selection of VI from a purely information theoretical perspective; in the next section, by describing its properties. The properties we focus on are mainly "geometric" properties; we show that VI is a metric, then we show this metric behaves in an intuitive and "desirable" way on clusterings that are related by elementary operations like splitting/merging of clusters.

We start by establishing how much information is there in each of the clusterings, and how much information one clustering gives about the other. For more details about the information theoretical concepts presented here, the reader is invited to consult [3].

Imagine the following game: if we were to pick a point of $D$, how much uncertainty is there about which cluster is it going to be in? Assuming that each point has an equal probability of being picked, it is easy to see that the probability of the outcome being in cluster $C_k$ equals

$$P(k) = \frac{n_k}{n} \tag{13}$$

Thus we have defined a discrete random variable taking $K$ values, that is uniquely associated to the clustering $\mathcal{C}$. The uncertainty in our game is equal

to the *entropy* of this random variable

$$H(\mathcal{C}) \;=\; -\sum_{k=1}^{K} P(k)\log P(k) \tag{14}$$

We call $H(\mathcal{C})$ the *entropy associated with clustering* $\mathcal{C}$. Entropy is always non-negative. It takes value 0 only when there is no uncertainty, namely when there is only one cluster. Entropy is measured in *bits*. The uncertainty of 1 bit corresponds to a clustering with $K = 2$ and $P(1) = P(2) = 0.5$. Note that the uncertainty does not depend on the number of points in $D$ but on the relative proportions of the clusters.

We now define the *mutual information* between two clusterings, i.e the information that one clustering has about the other. Denote by $P(k)$, $k = 1, \ldots K$ and $P'(k')$, $k' = 1, \ldots K'$ the random variables associated with the clusterings $\mathcal{C}$, $\mathcal{C}'$. Let $P(k, k')$ represent the probability that a point belongs to $C_k$ in clustering $\mathcal{C}$ and to $C'_{k'}$ in $\mathcal{C}'$, namely the joint distribution of the random variables associated with the two clusterings.

$$P(k, k') \;=\; \frac{|C_k \bigcap C'_{k'}|}{n} \tag{15}$$

We define $I(\mathcal{C}, \mathcal{C}')$ the mutual information between the clusterings $\mathcal{C}$, $\mathcal{C}'$ to be equal to the mutual information between the associated random variables

$$I(\mathcal{C},\, \mathcal{C}') \;=\; \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k, k')\log\frac{P(k, k')}{P(k)P'(k')} \tag{16}$$

Intuitively, we can think of $I(\mathcal{C}, \mathcal{C}')$ in the following way: We are given a random point in $D$. The uncertainty about its cluster in $\mathcal{C}'$ is measured by $H(\mathcal{C}')$. Suppose now that we are told which cluster the point belongs to in $\mathcal{C}$. How much does this knowledge reduce the uncertainty about $\mathcal{C}'$? This reduction in

uncertainty, averaged over all points, is equal to $I(\mathcal{C}, \mathcal{C}')$.

The mutual information between two random variables is always non-negative and symmetric.

$$I(\mathcal{C}, \mathcal{C}') \ = \ I(\mathcal{C}', \mathcal{C}) \ \geq \ 0 \tag{17}$$

Also, the mutual information can never exceed the total uncertainty in a clustering, so

$$I(\mathcal{C}, \mathcal{C}') \ \leq \ \min(H(\mathcal{C}), \ H(\mathcal{C}')) \tag{18}$$

Equality in the above formula occurs when one clustering completely determines the other. For example, if $\mathcal{C}'$ is obtained from $\mathcal{C}$ by merging two or more clusters, then

$$I(\mathcal{C}, \mathcal{C}') \ = \ H(\mathcal{C}') \ < \ H(\mathcal{C})$$

When the two clusterings are equal, and only then, we have

$$I(\mathcal{C}, \mathcal{C}') \ = \ H(\mathcal{C}') \ = \ H(\mathcal{C})$$

We propose to use as a comparison criterion for two clusterings $\mathcal{C}, \mathcal{C}'$ the quantity

$$VI(\mathcal{C}, \mathcal{C}') \ = \ H(\mathcal{C}) + H(\mathcal{C}') - 2I(\mathcal{C}, \mathcal{C}') \tag{19}$$

At a closer examination, this is the sum of two positive terms

$$VI(\mathcal{C}, \mathcal{C}') \ = \ [H(\mathcal{C}) - I(\mathcal{C}, \mathcal{C}')] + [H(\mathcal{C}') - I(\mathcal{C}, \mathcal{C}')] \tag{20}$$

FIGURE 2 GOES HERE

Fig. 2. The variation of information (represented by the sum of the shaded areas) and related quantities.

By analogy with the total variation of a function, we call it *variation of information* between the two clusterings. The two terms represent the conditional entropies $H(\mathcal{C}|\mathcal{C}'), H(\mathcal{C}'|\mathcal{C})$. The first term measures the amount of information about $\mathcal{C}$ that we loose, while the second measures the amount of information about $\mathcal{C}'$ that we have to gain, when going from clustering $\mathcal{C}$ to clustering $\mathcal{C}'$.

From the above considerations it follows that an equivalent expression for the variation of information (VI) is

$$VI(\mathcal{C},\mathcal{C}') \;=\; H(\mathcal{C}|\mathcal{C}') + H(\mathcal{C}'|\mathcal{C}) \tag{21}$$

Noting that

$$I(\mathcal{C},\mathcal{C}') \;=\; H(\mathcal{C}) + H(\mathcal{C}') - H(\mathcal{C},\mathcal{C}')$$

where $H(\mathcal{C},\mathcal{C}')$ is the entropy of $P(k,k')$, or the *joint entropy* of the two clusterings [3], we obtain a third equivalent expression for the variation of information

$$VI(\mathcal{C},\mathcal{C}') \;=\; 2H(\mathcal{C},\mathcal{C}') - H(\mathcal{C}) - H(\mathcal{C}') \tag{22}$$

## 4 Properties of the variation of information

We now list some basic properties of the variation of information with the goal of better understanding the structure it engenders on the set of all clusterings.

These properties will also help us decide whether this comparison criterion is appropriate for the clustering problem at hand. Here we will not be focusing on a specific application, but rather we will try to establish whether the properties are "reasonable" and in agreement with the general intuition of what "more different" and "less different" should mean for two clusterings of a set.

Most of the properties below have elementary proofs that are left as an exercise to the reader. The other proofs are given in the appendix.

### *4.1 The variation of information is a metric*

**Property 1** *The variation of information satisfies the metric axioms:*

**Non-negativity.** $VI(\mathcal{C},\mathcal{C}')$ *is always non-negative and* $VI(\mathcal{C},\mathcal{C}') = 0$ *if and only if* $\mathcal{C} = \mathcal{C}'$.

**Symmetry.** $VI(\mathcal{C},\mathcal{C}') = VI(\mathcal{C}',\mathcal{C})$

**Triangle inequality.** *For any 3 clusterings* $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ *of* $D$

$$VI(\mathcal{C}_1,\mathcal{C}_2) + VI(\mathcal{C}_2,\mathcal{C}_3) \geq VI(\mathcal{C}_1,\mathcal{C}_3) \tag{23}$$

Hence the VI is a *metric* (or *distance*) on clusterings. The space of all clusterings is finite, so this metric is necessarily bounded. A comparison criterion that is a metric has several important advantages. The properties of a metric – mainly the symmetry and the triangle inequality – make the criterion more understandable. Human intuition is more at ease with a metric than with an arbitrary function of two variables.

Second, the triangle inequality tells us that if two elements of a metric space (i.e clusterings) are close to a third they cannot be too far apart from each

other. This property is extremely useful in designing efficient data structures and algorithms. With a metric, one can move from simply comparing two clusterings to analyzing the structure of large sets of clusterings. For example, one can design algorithms á la K-means [9] that cluster a set of clusterings, one can construct ball trees of clusterings for efficient retrieval, or one can estimate the speed at which a search algorithm (e.g simulated annealing type algorithms) moves away from its initial point.

### *4.2 Upper bounds for VI*

As mentioned before, the VI metric is necessarily bounded, since there are only a finite number of clusterings of any data set $D$. The following set of properties give some intuition of scale in this metric space.

**Property 2** $n$**-invariance**. *The value of* $VI(\mathcal{C}, \mathcal{C}')$ *depends only on the relative sizes of the clusters. It does not directly depend on the number of points in the data set.*

**Property 3** *The following bound is attained for all* $n$*.*

$$VI(\mathcal{C}, \mathcal{C}') \leq \log n \tag{24}$$

For example, $\mathcal{C} = \{\{1\}, \{2\}, \{3\}, \ldots \{n\}\}$ and $\mathcal{C}' = \{D\}$ always achieve $VI(\mathcal{C}, \mathcal{C}') = \log n$.

We have said before that the VI distance does not depend on $n$. The bound in the above inequality however depends on $n$. This does not show a contradiction, but merely the fact that with more data points more clusterings are possible. For example, if two data sets $D_1, D_2$ have respectively $n_1, n_2$ points,

with $n_1 < n_2$ then no clustering of $D_1$ will have more than $n_1$ clusters, while for the set $D_2$ there can be clusterings with $K > n_1$ clusters.

If the number of clusters is bounded by a constant $K^*$ we can derive a bound that is dependent on $K^*$ only.

**Property 4** *If $\mathcal{C}$ and $\mathcal{C}'$ have at most $K^*$ clusters each, with $K^* \leq \sqrt{n}$, then $VI(\mathcal{C}, \mathcal{C}') \leq 2 \log K^*$.*

For any fixed $K^*$ the bound is approached arbitrarily closely in the limit of large $n$ and is attained in every case where $n$ is an exact multiple of $(K^*)^2$. This shows that for large enough $n$, clusterings of different data sets, with different numbers of data points, but with bounded numbers of clusters are really on the same scale in the metric VI.

The above consequence is extremely important if the goal is to compare clustering algorithms instead of clusterings of one data set only. The previous three properties imply that, everything else being equal, distances obtained from data sets of different sizes are comparable. For example, if one ran a clustering algorithm with the same parameters on 3 data sets produced by the same generative process, then one could compare the clusterings obtained by the algorithm with the gold standard for each of the 3 data sets and average the resulting 3 distances to obtain the average "error" of the algorithm. Other less restrictive comparisons are also possible and are being often done in practice, but their results should be regarded with caution. To summarize, if it makes sense to consider the clustering problems on two data sets as equivalent, then there is nothing intrinsic to the VI to prevent one from comparing, adding, subtracting VI distances across clustering spaces independently of the sizes of the underlying data sets.

Fig. 3. Two maximally separated clusterings $\mathcal{C}$ and $\mathcal{C}'$, having each $K = 3$ clusters, and their meet $\mathcal{C} \times \mathcal{C}'$, having 9 clusters.

FIGURE 3 GOES HERE

### *4.3 The local neighborhood*

A consequence of having a bounded metric with a known bound is that we can define $\epsilon$-radius balls around any clustering. The upper bound on the metric gives an absolute upper bound on $\epsilon$. Let us now look at the range of small $\epsilon$ values. Since the space is discrete, for fixed $n$, the balls with radius below a certain limit will contain only one clustering, the one around which they are centered. The following properties give the distances at which the nearest neighbors of a clustering $\mathcal{C}$ will lie. Note that these distances will always depend on $\mathcal{C}$.

In addition, we will also be concerned with characterizing the nearest clusterings to a given one. Or in other words, we will investigate the question: what changes to a clustering are small according to the VI distance?

**Property 5 Splitting a cluster**. *Assume $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting $C_k$ into clusters $C'_{k_1}, \ldots C'_{k_m}$. The cluster probabilities in $\mathcal{C}'$ are*

$$P'(k') \;=\; \begin{cases} P(k') & \text{if } C'_{k'} \in \mathcal{C} \\ P(k'|k)P(k) & \text{if } C'_{k'} \;\subseteq\; C_k \;\in\; \mathcal{C} \end{cases} \tag{25}$$

*In the above $P(k'|k)$ for $k' \in \{k_1, \ldots k_m\}$ is*

$$P(k_l|k) \;=\; \frac{|C'_{k_l}|}{|C_k|} \tag{26}$$

*and its entropy, representing the uncertainty associated with splitting $C_k$, is*

$$H_{|k} \;=\; -\sum_{l} P(k_l|k) \log P(k_l|k)$$

*Then,*

$$VI(\mathcal{C}, \mathcal{C}') = P(k) H_{|k} \tag{27}$$

The same value is obtained when performing the reverse operation, i.e when a set of clusters is merged into a single one. Equation (27) shows that the distance achieved by splitting a cluster is proportional to the relative size of the cluster times the entropy of the split. Hence, splitting (or merging) smaller clusters has less impact on the VI then splitting or merging larger ones. Note also that the variation of information at splitting or merging a cluster is independent of anything outside the cluster involved. This is a desirable property; things that are equal in two clusterings should not be affecting the distance between them.

Following from Property 5 we have that: (1) if $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting $C_k$ into $q$ equal clusters, then $VI(\mathcal{C}, \mathcal{C}') = P(k) \log q$, and (2) if $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting one point off $C_k$ and making it into a new cluster, then

$$VI(\mathcal{C}, \mathcal{C}') \;=\; \frac{1}{n}[n_k \log n_k - (n_k - 1) \log(n_k - 1)] \tag{28}$$

Since splitting off one point represents the lowest entropy split for a given cluster, it follows that splitting one point off the smallest non-singleton cluster results in the nearest $\mathcal{C}'$ with $K' > K$ to a given $\mathcal{C}$. This suggests that the nearest neighbors of a clustering $\mathcal{C}$ in the VI metric are clusterings obtained by splitting or merging small clusters in $\mathcal{C}$. In the following we prove that this

Fig. 4. Illustration of Property 6.

FIGURE 4 GOES HERE

is indeed so.

First some definitions. We shall say that a clustering $\mathcal{C}'$ *refines* another clustering $\mathcal{C}$ if for each cluster $C'_{k'} \in \mathcal{C}'$ there is a (unique) cluster $C_k \in \mathcal{C}$ so that $C'_{k'} \subseteq C_k$. In other words, a refinement $\mathcal{C}'$ is obtained by splitting some clusters of the original $\mathcal{C}$. If $\mathcal{C}'$ refines $\mathcal{C}$ it is easy to see that $K' \geq K$, with equality only if $\mathcal{C}' = \mathcal{C}$.

We define the *meet* of clusterings $\mathcal{C}$ and $\mathcal{C}'$ by

$$\mathcal{C} \times \mathcal{C}' \;=\; \{C_k \cap C'_{k'} \,|\, C_k \in \mathcal{C},\, C'_{k'} \in \mathcal{C}',\, C_k \cap C'_{k'} \neq \emptyset\}$$

Hence, the meet of two clusterings is the clustering formed from all the nonempty intersections of clusters from $\mathcal{C}$ with clusters from $\mathcal{C}'$. The meet $\mathcal{C} \times \mathcal{C}'$ contains all the information in $\mathcal{C}$ and $\mathcal{C}'$, i.e knowing a point's cluster in the meet uniquely determines its cluster in $\mathcal{C}$ and $\mathcal{C}'$. If $\mathcal{C}'$ is a refinement of $\mathcal{C}$, then $\mathcal{C} \times \mathcal{C}' = \mathcal{C}'$.

**Property 6 Collinearity of the meet**. *The triangle inequality holds with equality for two clusterings and their meet.*

$$VI(\mathcal{C}, \mathcal{C}') \;=\; VI(\mathcal{C}, \mathcal{C} \times \mathcal{C}') \;+\; VI(\mathcal{C}', \mathcal{C} \times \mathcal{C}') \tag{29}$$

The proof, elementary, is given in the appendix.

Thus, the meet of two clusterings is "collinear" with and "in between" the clusterings in this metric space, as depicted in figure 4. Finally, this leads us to the following property, which implies that the nearest neighbor of any

clustering $\mathcal{C}$ is either a refinement of $\mathcal{C}$ or a clustering whose refinement is $\mathcal{C}$.

**Property 7** *For any two clusterings we have*

$$VI(\mathcal{C},\mathcal{C}') \;\geq\; VI(\mathcal{C},\mathcal{C}\times\mathcal{C}') \tag{30}$$

*with equality only if* $\mathcal{C}' = \mathcal{C}\times\mathcal{C}'$.

From the above, we conclude that the nearest neighbor of $\mathcal{C}$, with $K' < K$ is obtained by merging the two smallest clusters in $\mathcal{C}$. We now have, due to equation (28) and property 7, a lower bound on the distance between a clustering $\mathcal{C}$ and any other clustering of the same data set. The lower bound depends on $\mathcal{C}$. Taking its minimum for all clusterings, which is attained when two singleton clusters are merged (or conversely, a cluster consisting of two points is split) we obtain

$$VI(\mathcal{C},\mathcal{C}') \;\geq\; \frac{2}{n} \quad \text{for all } \mathcal{C}\neq\mathcal{C}'$$

Equation (4.3) implies that the minimum distance between two clusterings decreases when the total number of points increases. In other words, the space of clusterings has not only a larger diameter for larger $n$ but it also has finer granularity. This is natural, since a larger $n$ allows clusterings not possible with smaller $n$'s. If we multiply $n$ by an integer $m$, obtaining $n' = m\times n$ and a new data set $D'$ that has $m$ points for each point of $D$, then it is easy to see that all the clusterings of $D$ are possible in $D'$ and that their respective distances in $D$ are preserved by the metric in $D'$. In addition, $D'$ will have clusterings not possible in $D$, that will be interspersed between the clusterings from $D$.

Fig. 5. Illustration of linearity. If $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$ and $\mathcal{C}' = \mathcal{C}'_1 \cup \mathcal{C}'_2$ then $VI(\mathcal{C}, \mathcal{C}') = \frac{n_1}{n_1+n_2} VI(\mathcal{C}_1, \mathcal{C}'_1) + \frac{n_2}{n_1+n_2} VI(\mathcal{C}_2, \mathcal{C}'_2)$.

FIGURE 5 GOES HERE

### *4.4 Convex additivity*

Looking at property 5 (splitting a cluster) from a different angle we can derive another interesting property of the variation of information.

**Property 8 Additivity of composition.** *Let $\mathcal{C} = \{C_1, \ldots C_K\}$ be a clustering and $\mathcal{C}'$, $\mathcal{C}''$ be two refinements of $\mathcal{C}$. Denote by $\mathcal{C}'_k$ ($\mathcal{C}''_k$) the partitioning induced by $\mathcal{C}'$ (respectively $\mathcal{C}''$) on $C_k$. Let $P(k)$ represent the proportion of data points that belong to cluster $C_k$. Then*

$$VI(\mathcal{C}', \mathcal{C}'') = \sum_{k=1}^{K} P(k) VI(\mathcal{C}'_k, \mathcal{C}''_k) \tag{31}$$

This property is illustrated in figure 5 for $K = 2$. The property can be interpreted in a way reminiscent of hierarchical clusterings. If two hierarchical clusterings have exactly two levels and they coincide on the higher level but differ on the lower level, then the VI distance between the two clusterings (regarded as flat clusterings) is a weighted sum of the VI distances between the second level partitions of each of the common first level clusters.

Property 8 can be seen in yet another way. If two disjoint clustered data sets are merged, they induce a clustering on their union. If there are two ways of clustering each of the data sets, the VI distance between any two induced clusterings of the union is a linear combination of the VI distances at the level of the component data sets.

## 5 Comparing VI with other criteria – scaling issues

Here we consider some of the other indices and metrics for comparing clusterings, and examine whether they can be made invariant with $n$ (of the criteria discussed in section 2 only the $\mathcal{H}$ and $\mathcal{L}$ criteria are). We compare the invariant versions that we obtain in order to better the understanding of these criteria. We give invariance with $n$ particular attention because, in any situation where comparisons are not restricted to a single data set, a criterion that is not $n$-invariant would have little value without being accompanied by the corresponding $n$.

The Rand, adjusted Rand, Fowlkes-Mallows, Jaccard, and Wallace indices are asymptotically $n$-invariant in the limit of large $n$. For finite values of $n$ the dependence on $n$ is weak. It is also non-linear, and we don't see a natural way of making these criteria exactly $n$-invariant.

A more interesting case is represented by the two metrics: the Mirkin metric $\mathcal{M}$, which is related to the Rand index and thus to counting pairs, and the van Dongen metric $\mathcal{D}$ based on set matching. Their dependence on $n$ is strong: both metrics grow unboundedly with $n$. However, both metrics can be easily scaled to become $n$-invariant. We denote the $n$-invariant versions of $\mathcal{D}$, $\mathcal{M}$ by $\mathcal{D}_{inv}$, $M_{inv}$.

$$\mathcal{D}_{inv}(\mathcal{C},\mathcal{C}') = \frac{\mathcal{D}(\mathcal{C},\mathcal{C}')}{2n}$$
$$\mathcal{M}_{inv}(\mathcal{C},\mathcal{C}') = \frac{\mathcal{M}(\mathcal{C},\mathcal{C}')}{n^2}$$

Since the Mirkin distance is related to the Rand index, by inspecting (9) we see that the Rand index is asymptotically equivalent to an $n$-invariant metric.

It is instructive to compare the behavior of the three invariant metrics $VI$, $\mathcal{M}_{inv}$, $\mathcal{D}_{inv}$ and of the widely used adjusted Rand index $\mathcal{AR}$ for two clusterings with $K$ clusters that are maximally separated under the VI distance. Such a situation is depicted in figure 3. The two clusterings have $n_k = n'_k = n/K$ and $n_{kk'} = n/K^2$ for all $k, k' = 1, \ldots K$. It is assumed that $n$ is a multiple of $K^2$ for simplicity. As random variables, these two clusterings are uniformly distributed and mutually independent. This pair of clusterings is also maximizing $\mathcal{D}_{inv}$ under the constraint that $K = K'$ but not $\mathcal{M}_{inv}$ (and consequently also not $\mathcal{AR}$).

We compute now the values $VI^0, \mathcal{D}^0_{inv}, \mathcal{M}^0_{inv}$ and $\mathcal{AR}^0$ for this particular pair, as a function of $K$.

$$
\begin{aligned}
VI^0 &= 2 \log K \\
\mathcal{D}^0_{inv} &= 1 - \frac{1}{K} \\
\mathcal{M}^0_{inv} &= \frac{2}{K} - \frac{1}{K^2} \\
\mathcal{AR}^0 &= -\frac{K-1}{n-K} \;\longrightarrow\; 0 \text{ asymptotically for } n \longrightarrow \infty
\end{aligned}
$$

It follows that while the VI distance grows logarithmically with $K$, the other two metrics have values bounded between 0 and 1 for any value of $K$. The $\mathcal{D}_{inv}$ metric grows with $K$ toward the upper bound of 1, while the $\mathcal{M}_{inv}$ metric *decreases* toward 0 approximately as $1/K$. As for the adjusted Rand index $\mathcal{AR}$, note that the above is not its minimum values, and that the index can take negative values for other pairs of clusterings.

In this context, note that it would be easy to normalize the variation of information by $\log n$ in order to obtain a distance that varies between 0 and 1.

$$\mathcal{V}(\mathcal{C}, \mathcal{C}') \;=\; \frac{1}{\log n} VI(\mathcal{C}, \mathcal{C}')$$

This is possible and convenient if we limit our comparison to one data set only. Normalizing by $\log n$ is however not recommended if we are to make comparisons between distances obtained on different data sets.

Another possibility is to normalize by the upper bound $2 \log K^*$ when the number of clusters is bounded by the same constant $K^*$ in all experiments.

$$\mathcal{V}_{K^*}(\mathcal{C}, \mathcal{C}') \;=\; \frac{1}{2 \log K^*} VI(\mathcal{C}, \mathcal{C}')$$

Such a normalization will preserve comparability of distances across data sets, independently of the number of data points in each set. Therefore it can be considered as a further simplification of the criterion. This simplification however has a drawback: while VI is always measured in bits, $\mathcal{V}_{K^*}$ will be measured in arbitrary units. Values of $\mathcal{V}_{K^*}$ obtained in different experiments, say by different authors using different values for $K^*$ will not be comparable without knowing the respective $K^*$ and renormalizing.

Hence, the adjusted forms of the VI distance lose one of the most valuable properties of the VI; that is, comparability across many different conditions.

## 6 Comparing VI with other criteria – Convex additivity and locality

Now we compare the scaled metrics with the VI distance from the point of view of convex additivity. The following proposition can be easily proved.

**Property 9 Convex additivity of composition for $\mathcal{D}_{inv}$, $\mathcal{M}_{inv}$.** *Let $\mathcal{C} = \{C_1, \dots C_K\}$ be a clustering and $\mathcal{C}', \mathcal{C}''$ be two refinements of $\mathcal{C}$. Denote by $\mathcal{C}'_k$ ($\mathcal{C}''_k$) the partitioning induced by $\mathcal{C}'$ (respectively $\mathcal{C}''$) on $C_k$. Let $n_k$ represent the number of data points that belong to cluster $C_k$. Then*

$$\mathcal{D}_{inv}(\mathcal{C}', \mathcal{C}'') = \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{D}_{inv}(\mathcal{C}'_k, \mathcal{C}''_k)$$
$$\mathcal{M}_{inv}(\mathcal{C}', \mathcal{C}'') = \sum_{k=1}^{K} \frac{n_k^2}{n^2} \mathcal{M}_{inv}(\mathcal{C}'_k, \mathcal{C}''_k)$$

Hence, the $\mathcal{D}_{inv}$ metric behaves like the VI metric in that the resulting distance is a convex combination of the distances between the subclusterings. The $\mathcal{M}_{inv}$ metric is linear too, but the coefficients depend quadratically on $n_k/n$ so that the resulting distance is smaller than the convex combinations of distances between subclusterings. This is in agreement with equation (32) showing that the Mirkin metric can decreases rapidly with the number of clusters. Note also that the unscaled versions of $\mathcal{D}$, $\mathcal{M}$ are additive.

Additivity for a metric entails the following property, called *locality*: If $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting one cluster, then the distance between $\mathcal{C}$ and $\mathcal{C}'$ depends only on the cluster undergoing the split. Metrics that are additive are also *local*. For example, for the Mirkin metric in the case of splitting cluster

$C_k$ into $C_k^1, C_k^2$, locality is expressed as

$$\mathcal{M}_{inv}(\mathcal{C}, \mathcal{C}') = \frac{n_k^2}{n^2} \mathcal{M}_{inv}(\{C_k\}, \{C_k^1, C_k^2\}) \tag{32}$$

The r.h.s of the above formula depends only on quantities related to $C_k$ and its split. It is invariant to the configuration of the other clusters in the partition. Locality for the VI distance is reflected by property 5.

The VI distance as well as the $\mathcal{D}$ and $\mathcal{M}$ metrics and their $n$-invariant versions are local. It can be easily shown that the Rand and the classification error $\mathcal{H}$ indices are also local. The Larsen index is not local. For the split $C_k \rightarrow C_k^1, C_k^2$ we have

$$\mathcal{L}(\mathcal{C}, \mathcal{C}') = \frac{1}{K} \left[ K - 1 + 2 \max_{i=1,2} \frac{|C_k^i|}{|C_k| + |C_k^i|} \right]$$
$$\mathcal{L}(\mathcal{C}', \mathcal{C}) = \frac{1}{K+1} \left[ K - 1 + 2 \sum_{i=1,2} \frac{|C_k^i|}{|C_k| + |C_k^i|} \right]$$

The above expressions depend on $K$ so are not local.

For the Fowlkes-Mallows and Jaccard indices we obtain:

$$\mathcal{F}(\mathcal{C}, \mathcal{C}') = \sqrt{1 - \frac{2|C_k^1||C_k^2|}{n_k^2 - n + \sum_{l \neq k} n_l^2}}$$
$$\mathcal{J}(\mathcal{C}, \mathcal{C}') = 1 - \frac{2|C_k^1||C_k^2|}{n_k^2 - n + \sum_{l \neq k} n_l^2}$$

The common fraction on the r.h.s of both expressions depends on $\sum_{l \neq k} n_l^2$ so the indices depend strongly on how the rest of the points are clustered.

We will analyze now the asymptotic, $n$-invariant expression of the adjusted Rand index. This is

$$\mathcal{AR}(\mathcal{C},\mathcal{C}') = \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K'}P(k,k')^2 - \left(\sum_{k=1}^{K}P(k)^2\right)\left(\sum_{k'=1}^{K'}P'(k')^2\right)}{\frac{\sum_{k=1}^{K}P(k)^2+\sum_{k'=1}^{K'}P'(k')^2}{2} - \left(\sum_{k=1}^{K}P(k)^2\right)\left(\sum_{k'=1}^{K'}P'(k')^2\right)} \quad (33)$$

We assume as before that $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting $C_k$, into $C_k^1$, $C_k^2$ with respective cluster probabilities $P_k^1$, $P_k^2$, $P_k^1 + P_k^2 = P(k)$. Denote also $Q_k = (P_k^1)^2 + (P_k^2)^2$ and $S = \sum_{j\neq k} P(j)^2$ After calculations we obtain

$$\mathcal{AR}(\mathcal{C},\mathcal{C}') = 1 - \frac{P_k^1 P_k^2}{\frac{Q_k+P(k)^2}{2} - Q_k P(k)^2 - [Q_k + P(k)^2 - 1]S - S^2} \quad (34)$$

The denominator of the fraction above depends, through the expression $S = \sum_{j\neq k} n_j^2/n^2$, on the way the rest of the data is partitioned. Hence, the asymptotic expression of the $\mathcal{AR}$ index is not local. This implies (by contradiction) that the $\mathcal{AR}$ itself is not local.

Whether a criterion for comparing clusterings should be local or not depends ultimately on the specific requirements of the application. A priori, however, a local criterion is more intuitive and easier to understand.

## 7 The axioms of VI

Constructing a quantity axiomatically highlights which of its properties, taken together, represent defining properties for it, from which all others follow. This is what we do now with the VI.

**Property 10** *The variation of information is the unique cluster comparison*

*criterion $d$ that satisfies the axioms:*

*A1* **Vertical collinearity** *Denote by $\hat{0}$ and $\hat{1}$ the unique clusterings having $K = n$ respectively $K = 1$ clusters. For any clustering $\mathcal{C}$*

$$d(\hat{1}, \mathcal{C}) + d(\mathcal{C}, \hat{0}) \;=\; d(\hat{1}, \hat{0})$$

*A2* **Horizontal collinearity** *For any two clusterings $\mathcal{C}, \mathcal{C}'$*

$$d(\mathcal{C}, \mathcal{C}') \;=\; d(\mathcal{C}, \mathcal{C} \times \mathcal{C}') + d(\mathcal{C}', \mathcal{C} \times \mathcal{C}')$$

*A3* **Convex additivity** *Let $\mathcal{C} = \{C_1, \ldots C_K\}$ be a clustering and $\mathcal{C}'$ be a refinement of $\mathcal{C}$. Denote by $\mathcal{C}'_k$ the partitioning induced by $\mathcal{C}'$ on $C_k$. Then*

$$d(\mathcal{C}, \mathcal{C}') \;=\; \sum_{k=1}^{K} \frac{n_k}{n} d(\hat{1}_{n_k}, \mathcal{C}'_k)$$

*A4* **Scale** *Denote by $\mathcal{C}_K^U$ the "uniform" clustering, i.e the clustering with $K$ equal clusters, $K \geq 1$. If $\mathcal{C}_K^U$ exists, then*

$$d(\hat{1}, \mathcal{C}_K^U) \;=\; \log K$$

The proof is constructive and is given in the appendix. Note that the axioms do not require $d$ to be a metric or to be $n$-invariant; these properties follow implicitly.

One can represent all clusterings of $D$ as the nodes of a graph; in this graph an edge between $\mathcal{C}, \mathcal{C}'$ will be present if $\mathcal{C}'$ is obtained by splitting a cluster of $\mathcal{C}$ into two parts. The set of all clusterings of $D$ forms a lattice and this graph is known as the *Hasse diagram*[3] of the *lattice of partitions* [17]. Conventionally $\hat{1}$ is represented at the top of the diagram and $\hat{0}$ at the bottom. Axioms A1

[3] The emphasized terms in this section represent standard lattice terminology. Their precise definitions can be found in e.g [17]

Fig. 6. The lattice of partitions of $D$ ={a,b,c,d}. Note that $\hat{1}$ ={{a,b,c,d}}, $\hat{0}$ ={{a},{b},{c},{d}}; the clusterings $\hat{1}$, {{a},{b, c, d}}, {{a},{b},{c, d}}, $\hat{0}$ are collinear according to A1; the clusterings {{a},{b, c, d}}, {{a},{b},{c, d}}, {{a,b},{c, d}} are collinear according to A2; and there are 3 straight lines from {{d},{a, b, c}} to $\hat{0}$.

FIGURE 6 GOES HERE

and A2 show that the VI metric is "aligned" with the lattice of partitions and embeds it into a metric space. Axiom A1 implies that the clusterings along a vertical *chain* of lattice edges are collinear under the VI metric; in addition, by A2, two vertical chains that *meet* in their lowest vertex also represent a "straight line" according to VI.

Note that, if $\mathcal{C}'$ is a refinement of $\mathcal{C}$, then each of the possible ways of subdividing $\mathcal{C}$ to obtain $\mathcal{C}'$ generates a straight line in the lattice. Unless $\mathcal{C}, \mathcal{C}'$ are connected by a single edge, there will be multiple "straight lines" between the two clusterings. Figure 6 illustrates these properties on a simple example.

Finally, axioms A3 and especially A4 set the scale of $d$. They are inspired by the postulates III and IV of entropy as given in [16]. Axiom A3 is a rewrite of property 5, which states that if one cluster is split, then the distance from the new clustering to the original equals the distance induced by the split scaled by the size of the cluster.

It is interesting to see what happens if these last axioms are changed. In other words, if one maintains that the distance has to be aligned with the lattice of partitions, but allows the scale to differ.

For example, one can compute the distance between $\hat{1}$ and $\mathcal{C}_K^U$ under the invariant Van Dongen and Mirkin metrics and plug it into A4. We have

$$\mathcal{D}_{inv}(\hat{1}, \mathcal{C}_K^U) = \frac{1}{2}\left(1 - \frac{1}{K}\right)$$
$$\mathcal{M}_{inv}(\hat{1}, \mathcal{C}_K^U) = 1 - \frac{1}{K}$$

**Property 11** *There is no cluster comparison criterion that satisfies axioms A1–A3 and*
*A4.M* $d(\hat{1}, \mathcal{C}_K^U) = 1 - 1/K$.

We can however recover the Mirkin metric by modifying the convex additivity axiom in accordance to equation (32).

**Property 12** *The unique cluster comparison criterion* $d$ *that satisfies axioms A1, A2, A4.M and*
*A3.M Let* $\mathcal{C}'$ *be a refinement of* $\mathcal{C}$ *and denote by* $\mathcal{C}'_k$ *the clustering induced by* $\mathcal{C}'$ *on* $C_k \in \mathcal{C}$. *Then*

$$d(\mathcal{C}, \mathcal{C}') = \sum_{k=1}^{K} \frac{n_k^2}{n^2} d(\hat{1}_{n_k}, \mathcal{C}'_k)$$

*is the invariant Mirkin metric* $\mathcal{M}_{inv}$.

Thus, the invariant Mirkin metric is also aligned with the lattice of partitions. It is worth recalling that all these properties of the Mirkin metric are readily translated into similar properties of the unadjusted Rand index. From the above, it follows that the Rand index is approximately aligned with the partitions lattice as well.

In [14], the Mirkin metric is constructed from a different set of axioms, which we transcribe here in terms of the $\mathcal{M}_{inv}$ metric.

M1. $d$ is a metric

M2. $d(\hat{1}, \{\{i\}, D \setminus \{i\}\}) = 2(n-1)/n^2$.

M3. Let $\mathcal{C}' = \{C_1, \ldots, C_m, C'_{m+1}, \ldots, C'_{K'}\}$ and $\mathcal{C}'' = \{C_1, \ldots, C_m, C''_{m+1}, \ldots, C''_{K''}\}$ be two clusterings which have some clusters in common. Denote by $E = D \setminus (\bigcup_{k=1}^{m} C_k)$ and by $\mathcal{C}'_E$, $(\mathcal{C}''_E)$ the clustering that $\mathcal{C}'$, $(\mathcal{C}'')$ induces on $E$. Then $d(\mathcal{C}', \mathcal{C}'') = d(\mathcal{C}'_E, \mathcal{C}''_E)$.

M4. $d(\hat{1}, \hat{0}) = (n-1)/n$.

One can see that the last axiom is a weaker form of the scale axiom A4.M, and that axiom M3 is a stronger form of locality, most similar to our A3 axiom. The remaining axioms M1 and M2 are entirely different from ours; in particular, M1 requires that $d$ be a metric.

Property 11 prompts the question: what kind of scalings in A4 are compatible with A1–A3? To answer this question, we change A4 to the weaker A4.H:

*A4.H* $d(\hat{1}, \mathcal{C}_K^U) = h(K)$ *where* $h$ *is a non-decreasing function of* $K$, $K \geq 1$.

The result below shows that A4 is essentially superfluous.

**Property 13** *Any clustering comparison criterion satisfying A1–A3 and A4.H is identical to VI up to a multiplicative constant.*

In other words, the VI is the only sensible (i.e $n$-invariant, with $d(\hat{1}, \mathcal{C}_K^U)$ non-decreasing), criterion that is convexely additive and aligned to the lattice of partitions.

## 8 Concluding remarks

This paper has presented a new criterion for comparing two clusterings of a data set, that is derived from information theoretic principles.

The criterion is more discriminative than the previously introduced criteria that are based on set matching. In particular, for the example in figure 1 we have $VI(\mathcal{C}, \mathcal{C}'') = VI(\mathcal{C}, \mathcal{C}') + 2f \log(K-1)$ implying that $\mathcal{C}'$ is closer to $\mathcal{C}$ than $\mathcal{C}''$ for all $K > 2$.

In contrast to the comparison criteria based on counting pairs, the variation of information is not directly concerned with relationships between pairs of points, or with triples like [6]. One could say that the variation of information is based on the relationship between a point and its cluster in each of the two clusterings that are compared. This is neither a direct advantage, nor a disadvantage w.r.t the criteria based on pair counts.

The vast literature on the subject suggests that criteria like $\mathcal{R}$, $\mathcal{F}$, $\mathcal{K}$, $\mathcal{J}$ need to be shifted and rescaled in order to allow their values to be compared. However, the existing rescaling methods are based on a null model which, although reasonable, is nevertheless artificial. By contrast, the variation of information makes no assumptions about how the clusterings may be generated and requires no rescaling to compare values of $VI(\mathcal{C}, \mathcal{C}')$ for arbitrary pairs of clusterings of a data set.

Moreover, the variation of information does not directly depend on the number of data points in the set. This gives a much stronger ground for comparisons across data sets, something we need to do if we want to compare clustering algorithms against each other. It may also pave the way to comparing the "finite sample" results of standard clustering with "population values" based on theoretically infinite data.

As $K$ grows, the VI distance between two clusterings can grow as large as $2 \log K$. This sets the VI distance apart from all other indices and met-

rics discussed here. The scaled metrics $\mathcal{M}_{inv}$, $\mathcal{D}_{inv}$ as well as the indices $\mathcal{R}$, $\mathcal{F}$, $\mathcal{J}$, $\mathcal{W}$, $\mathcal{H}$ are bounded between 0 and 1. Hence they carry the implicit assumption that clusterings can only get negligibly more diverse if at all as the number of clusters increases. Whether a bounded or unbounded criterion for comparing clusterings is better depends on the clustering application at hand. This paper's aim in this respect is to underscore the possible choices.

In the practice of comparing clusterings, one deals more often with clusterings that are close to each other than with clusterings that are maximally apart. For example, one often needs to compare partitions obtained by several clustering algorithms to a gold standard. It is reasonable to expect that the clusterings so obtained are somewhat similar to each other. The results on locality and the local neighborhood in section 4.3 help one understand the behavior of VI in this context. Note for example that the fact that the maximum VI distance grows like $\log K$ does not affect the local properties of the variation of information.

It has been shown here that VI is a metric. This is extremely fortunate as it allows one to see past simple pairwise comparisons between clusterings into the global structure of the space of clusterings. A metric also entails the existence of local neighborhoods, and this in turn allows us to apply to clusterings a vast array of already existing algorithmic techniques. One could for example cluster a set of clusterings obtained by different algorithms. This has already been suggested as a tool for results summarization but so far no existent metric has been used for this problem.

Just as one cannot define a "best" clustering method out of context, one cannot define a criterion for comparing clusterings that fits every problem optimally.

This paper has strived to present a comprehensible picture of the properties of the VI criterion, in order to allow a potential user to make informed decisions.

Some of the properties of the variation of information are shared with other criteria. For example, the VI between two clusterings with $K$ clusters each cannot be higher than $2 \log K$. Hence, for $K$ small, no two clusterings can be too far apart. This behavior is reflected in both the Fowlkes-Mallows and the Jaccard indices. For two clusterings with $K = 2$ that are maximally separated under the VI metric $\mathcal{F} = 0.5$ and $\mathcal{J} = 0.13$; these values become $\mathcal{F} = 0.16$ and $\mathcal{J} = 0.02$ if $K = 5$. The effect is due to an intrinsic property of the space of partitions: for $K$ small, no matter how a set is partitioned, there will be large overlaps between clusters of different partitions.

The axiomatic approach has not only characterized the VI but has also highlighted interesting impossibility results pertaining to any clustering comparison criterion. Thus, the properties of variation of information presented here, beyond enabling us to understand the VI metric, represent a tool that will help us think about the space of clusterings in a precise way and will sharpen our intuitions about it.

**Acknowledgements**

This work was partially supported by NSF grant IIS-0313339. The anonymous reviewers are gratefully acknowledged for the additional references they provided, and for a remark which led to a simplification in the axiomatic characterization of VI.

**Proofs**

**Proof of Property 1** As non-negativity and symmetry are evident from the definition, here we focus on the triangle inequality. We prove that for any clusterings $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$

$$H_{2|1} + H_{3|2} - H_{3|1} \;\geq\; 0 \tag{35}$$

In the above we used the shorthand notation $H_{p|q}$ for $H(\mathcal{C}_p|\mathcal{C}_q)$.

$$\begin{aligned} H_{2|1} + H_{3|2} - H_{3|1} &\geq H_{2|1} + H_{3|2,1} - H_{3|1} && (36)\\ &= H_{2,3|1} - H_{3|1} && (37)\\ &\geq 0 \end{aligned}$$

The first inequality is true because conditioning always decreases entropy, the second because the joint entropy is always larger than the marginal entropy. For more detail, see [3].

From (35) by swapping indices 1 and 3 and then adding the two inequalities, we obtain

$$H_{2|1} + H_{1|2} + H_{3|2} + H_{2|3} - H_{3|1} - H_{1|3} \;\geq\; 0 \tag{38}$$

which is the triangle inequality for the variation of information.

**Proof of Property 6**

First we prove that

$$H(\mathcal{C}|\mathcal{C}') \;=\; H(\mathcal{C} \times \mathcal{C}'|\mathcal{C}') \tag{39}$$

Let $C_l^*$ be a cluster in $\mathcal{C} \times \mathcal{C}'$ and let $C^*_{l_{kk'}} = C_k \cap C'_{k'}$ if it is non-empty. Note that for all clusters $C_k$ that intersect $C'_{k'}$

$$P(k|k') \;=\; P(l_{kk'}|k').$$

Then,

$$\begin{aligned}
H(\mathcal{C} \times \mathcal{C}'|\mathcal{C}') &= -\sum_{k'} P'(k') \sum_{l} P(l|k') \log P(l|k') \\
&= -\sum_{k'} P'(k') \left[ \sum_{C_l^* \subseteq C'_{k'}} P(l|k') \log P(l|k') + \sum_{C_l^* \not\subseteq C'_{k'}} 0 \right] \\
&= -\sum_{k'} P'(k') \sum_{C_k \cap C'_{k'} \neq \emptyset} P(k|k') \log P(k|k') \\
&= H(\mathcal{C}|\mathcal{C}')
\end{aligned}$$

Noting that $H(\mathcal{C}|\mathcal{C} \times \mathcal{C}') = H(\mathcal{C}'|\mathcal{C} \times \mathcal{C}') = 0$ we get

$$\begin{aligned}
VI(\mathcal{C}, \mathcal{C} \times \mathcal{C}') \;+\; VI(\mathcal{C}', \mathcal{C} \times \mathcal{C}') &= H(\mathcal{C} \times \mathcal{C}'|\mathcal{C}) + H(\mathcal{C} \times \mathcal{C}'|\mathcal{C}') \\
&= H(\mathcal{C}'|\mathcal{C}) + H(\mathcal{C}|\mathcal{C}') \\
&= VI(\mathcal{C}, \mathcal{C}')
\end{aligned}$$

**Proof of Property 10** From A3 and A4 we have that

$$\begin{aligned}
d(\mathcal{C}, \hat{0}) &= \sum_k \frac{n_k}{n} d(\hat{1}, \mathcal{C}^U_{n_k}) && (40) \\
&= \sum_k \frac{n_k}{n} \log n_k && (41) \\
&= \sum_k \frac{n_k}{n} (\log \frac{n_k}{n} + \log n) && (42) \\
&= \log n - H(\mathcal{C}) && (43)
\end{aligned}$$

From A1 we get $d(\hat{1}, \mathcal{C}) \;=\; \log n - d(\mathcal{C}, \hat{0}) \;=\; H(\mathcal{C})$. For any two clusterings $\mathcal{C}, \mathcal{C}'$ define by $\mathcal{C}_k$ the clustering induced by $\mathcal{C}'$ on $C_k \in \mathcal{C}$.

$$d(\mathcal{C}, \mathcal{C} \times \mathcal{C}') = \sum_{k=1}^{K} \frac{n_k}{n} d(\hat{1}, \mathcal{C}_k) \qquad (44)$$

$$= \sum_{k=1}^{K} \frac{n_k}{n} H(\mathcal{C}_k) \tag{45}$$

$$= H(\mathcal{C}|\mathcal{C}') \tag{46}$$

Therefore, by A2, $d(\mathcal{C}, \mathcal{C}') = H(\mathcal{C}|\mathcal{C}') + H(\mathcal{C}'|\mathcal{C})$, Q.E.D

**Proof of Property 11**

$$d(\hat{0}, \mathcal{C}) = \sum_{k} \frac{n_k}{n} d(\hat{1}, \mathcal{C}^U_{n_k}) \tag{47}$$

$$= \sum_{k} \frac{n_k}{n} \left(1 - \frac{1}{n_k}\right) \tag{48}$$

$$= 1 - \frac{K}{n} \tag{49}$$

Therefore $d(\hat{1}, \mathcal{C}) = (1 - 1/n) - (1 - K/n) = (K-1)/n$ if $|\mathcal{C}| = K$. This contradicts A4 according to which $d(\hat{1}, \mathcal{C}^U_K) = (K-1)/K$.

**Proof of Property 12** This proof follows the same steps as the proof of Property 10 and is therefore left to the reader.

**Proof of Property** 13 We have consecutively:

$$d(\hat{1}, \hat{0}) = h(n) \quad \text{by A5.H} \tag{50}$$

$$d(\hat{1}, \mathcal{C}) = h(n) - d(\mathcal{C}, \hat{0}) \quad \text{by A1} \tag{51}$$

$$d(\mathcal{C}, \hat{0}) = \sum_{k} \frac{n_k}{n} h(n_k) \quad \text{by A3} \tag{52}$$

$$d(\hat{1}, \mathcal{C}^U_K) = h(n) - d(\mathcal{C}^U_K, \hat{0}) = h(n) - K\frac{1}{K} h(\frac{n}{K}) = h(n) - h(\frac{n}{K}) \tag{53}$$

Since $n/K = M$ is an integer, and recalling A4.H we can rewrite the last equality as

$$h(K) = h(KM) - h(M)$$

or equivalently

$$h(KM) \;=\; h(K) + h(M) \tag{54}$$

for any positive integers $K, M$. By lemma 14 below, this implies that $h(n) = C \log n$ for all $n = 1, 2, 3, \ldots$.

It follows that A1-A3 together with A4.H imply essentially the original A4 (up to the multiplicative constant $C$) and therefore $d$ cannot be but proportional to the VI.

**Lemma 14** *Let $h : \{1, 2, \ldots\} \to [0, \infty)$ be a non-decreasing function satisfying (54) for any positive integers $K, M$. Then $h(n) = C \log n$ for any $n$.*

**Proof** (Thanks to Pavel Krivitzky and Nilesh Dalvi for providing each a proof). Let $h(2) = C$. We prove that $h(n) = C \log n$. Let $a = \log(n^q) = q \log n$ with $q$ a large positive integer. Then

$$h(2^{\lfloor a \rfloor}) \leq\; h(2^a) \;=\; h(n^q) \;\leq h(2^{\lceil a \rceil}) \tag{55}$$

$$\lfloor a \rfloor C \quad \leq\; q h(n) \;\leq \quad \lceil a \rceil C \tag{56}$$

$$\frac{\lfloor a \rfloor}{a} \quad \leq\; \frac{h(n)}{C \log n} \;\leq \quad \frac{\lceil a \rceil}{a} \tag{57}$$

The middle term does not depend on $q$, while the left and right tend to 1 for $q$ increasing to infinity, which implies $h(n) = C \log n$.

## References


[1] Michael R. Anderberg. *Cluster analysis for applications.* Academic Press, New York, NY, 1973.

[2] Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17, 2002.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[4] 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.

[5] John A. Hartigan. *Clustering algorithms*. Wiley, New York, NY, 1975.

[6] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[7] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.

[8] B. Larsen and C. Aone. Fast and effective text mining using linear time document clustering. In *Proceedings of the conference on Knowledge Discovery and Data Mining*, pages 16–22, 1999.

[9] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[10] G. J. MacLachlan and K. E. Bashford. *Mixture models: Inference and applications to Clustering*. Marcel Dekker, NY, 1988.

[11] Marina Meilă. Comparing clusterings – an axiomatic view. In Stefan Wrobel and Luc De Raedt, editors, *Proceedings of the International Machine Learning Conference (ICML)*. ACM Press, 2005.

[12] Marina Meilă and David Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1/2):9–29, 2001.

[13] Boris G. Mirkin. *Mathematical classification and clustering*. Kluwer Academic Press, 1996.

[14] Boris G. Mirkin and L. B. Cherny. Measurement of the distance between distinct partitions of a finite set of objects. *Automation and Remote Control*, 31(5):786, 1970.

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

[16] Alfred Rènyi. *Probability theory*. North-Holland, 1970.

[17] Richard P. Stanley. *Enumerative Combinatorics*. Cambridge Unversity Press, 1997.

[18] Douglas L. Steinley. Properties of the Hubert-Arabie adjusted Rand index. *Psychological methods*, 9(3):386–396, 2004. simulations of some adjusted indices and of misclassification error.

[19] Stijn van Dongen. Performance criteria for graph clustering and Markov cluster experiments. Technical Report INS-R0012, Centrum voor Wiskunde en Informatica, 2000.

[20] David L. Wallace. Comment. *Journal of the American Statistical Association*, 78(383):569–576, 1983.