Comparing Clusterings

Marina Meilă

University of Washington
Box 354322
Seattle WA 98195-4322
mmp@stat.washington.edu

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 criterion makes no assumptions about how the clusterings were generated and applies to both soft and hard clusterings. The basic properties of VI are presented and discussed from the point of view of comparing clusterings. In particular, the VI is positive, symmetric and obeys the triangle inequality. Thus, surprisingly enough, it is a true metric on the space of clusterings.

Keywords: Clustering; Comparing partitions; Measures of agreement; Information theory; Mutual information

1 Introduction

This paper proposes a simple information theoretic criterion for comparing two clusterings. The concepts of entropy and information have proved themselves as useful 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 this quality of information theoretic concepts. 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, the paper 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 paper starts with presenting previously used comparison criteria (section 2). The variation of information is introduced in section 3 and its properties are presented in section 4. In section 5 the variation of information is compared with other metrics and criteria of similarity between clusterings.

2 Related work

A clustering $\mathcal{C}$ is a partition of a set of points, or *data set* D into sets $C_1, C_2, \ldots C_K$ called *clusters* such that $C_k \cap C_l = \emptyset$ 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 $n = \sum_{k=1}^{K} n_k$. 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 $[n_{kk'}]$. See [3] for details.

Wallace [12] proposed the two asymmetric criteria $\mathcal{W}_I, \mathcal{W}_{II}$ below.

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

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

$$\mathcal{F}(\mathcal{C},\mathcal{C}') \;=\; \sqrt{\mathcal{W}_I(\mathcal{C},\mathcal{C}')\mathcal{W}_{II}(\mathcal{C},\mathcal{C}')} \tag{2}$$

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" clusterings [3]. 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. The adjusted Rand index is a similar transformation introduced by [4] of Rand's [10] criterion

$$\mathcal{R}(\mathcal{C},\mathcal{C}') \;=\; \frac{N_{11}+N_{00}}{n(n-1)/2} \tag{3}$$

A problem with adjusted indices is that the baseline is an expectation under a null hypothesis. The null hypothesis is that a) the two clusterings are sampled independently, and b) the clusterings are sampled from the set of all partition pairs with fixed n_k, $n'_{k'}$ points in each cluster [3, 4]. In practice, the second assumption is normally violated. Many algorithms take a number of clusters K as input, but the numbers of points in each cluster are a result of the execution of the algorithm. In most exploratory data analysis situations, it is unnatural to assume that anyone can know exactly how many points are in each cluster. The problems listed above have been known in the statistical community for a long time; see for example [12].

On the other hand, the range of values of the unadjusted $\mathcal{F}$ and $\mathcal{R}$ varies sharply for values of K, K' smaller than $n/3$ making comparisons across different values of K, K' unreliable [3].

There are other criteria in the literature, to which the above discussion applies, such as the Jacard [1] index

$$\mathcal{J}(\mathcal{C}, \mathcal{C}') = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{4}$$

an improved version of the Rand index, and the Mirkin [9] metric

$$\mathcal{M}(\mathcal{C}, \mathcal{C}') = \sum_k n_k^2 + \sum_{k'} {n'_{k'}}^2 - 2 \sum_k \sum_{k'} n_{kk'}^2 \tag{5}$$

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 [9]. This metric can also be rewritten as

$$\mathcal{M}(\mathcal{C}, \mathcal{C}') = 2(N_{01} + N_{10}) = n(n-1)[1 - \mathcal{R}(\mathcal{C}, \mathcal{C}')] \tag{6}$$

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. Meilă and Heckerman [8] computed the criterion $\mathcal{H}$: First, each cluster of $\mathcal{C}$ is given a "best match" in $\mathcal{C}'$. This is done by scanning the elements $n_{kk'}$ of the contingency table in decreasing order. The largest of them, call it n_{ab}, entails a match between C_a and C'_b, the second largest not in row a or column b entails the second match, and so on until $\min(K, K')$ matches are made. Denote by $match(k)$ the index of the cluster $C'_{k'}$ in $\mathcal{C}'$ that matches cluster C_k. Then

$$\mathcal{H}(\mathcal{C}, \mathcal{C}') = \frac{1}{n} \sum_{k'=match(k)} n_{kk'} \tag{7}$$

The index is symmetric and takes value 1 for identical clusterings. Larsen et al., [5] use

$$\mathcal{L}(\mathcal{C}, \mathcal{C}') = \frac{1}{K} \sum_k \max_{k'} \frac{2 n_{kk'}}{n_k + n'_k} \tag{8}$$

This is an asymmetric criterion that is 1 when the clusterings are identical. A criterion that is a metric was introduced by van Dongen [11]

$$\mathcal{D}(\mathcal{C},\mathcal{C}') \;=\; 2n - \sum_{k} \max_{k'} n_{kk'} - \sum_{k'} \max_{k} n_{kk'} \tag{9}$$

All three above criteria 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. For example, 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, the criterion we propose for comparing two clusterings.

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 [2].

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

$$P(k) = \frac{n_k}{n} \tag{10}$$

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

$$H(\mathcal{C}) \;=\; -\sum_{k=1}^{K} P(k) \log P(k) \tag{11}$$

We call $H(\mathcal{C})$ the *entropy associated with clustering* $\mathcal{C}$. Entropy is always nonnegative. It takes value 0 only when there is no uncertainty, namely when there is only one cluster. 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.

$$P(k, k') \;=\; \frac{|C_k \bigcap C'_{k'}|}{n} \tag{12}$$

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

$$I(\mathcal{C}, \mathcal{C}') \;=\; \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P'(k')} \tag{13}$$

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.

$$I(\mathcal{C}, \mathcal{C}') \;=\; I(\mathcal{C}', \mathcal{C}) \;\geq\; 0 \tag{14}$$

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

$$I(\mathcal{C}, \mathcal{C}') \;\leq\; \min(H(\mathcal{C}),\; H(\mathcal{C}')) \tag{15}$$

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

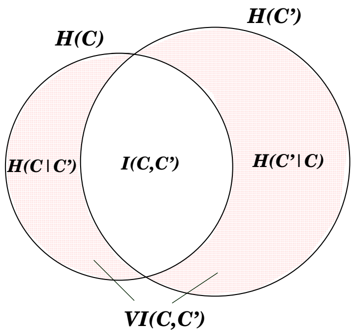
$$VI(\mathcal{C}, \mathcal{C}') \;=\; H(\mathcal{C}) + H(\mathcal{C}') - 2I(\mathcal{C}, \mathcal{C}') \tag{16}$$

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

$$VI(\mathcal{C}, \mathcal{C}') \;=\; [H(\mathcal{C}) - I(\mathcal{C}, \mathcal{C}')] + [H(\mathcal{C}') - I(\mathcal{C}, \mathcal{C}')] \tag{17}$$

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 $\mathcal{C}'$.

Fig. 1. The variation of information (represented by the sum of the shaded areas) and related quantities.



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

$$VI(\mathcal{C},\mathcal{C}') \;=\; H(\mathcal{C}|\mathcal{C}') + H(\mathcal{C}'|\mathcal{C}) \tag{18}$$

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 [2], we obtain a third equivalent expression for the variation of information

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

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 proofs for properties 1, 8 are given in the long version of the paper [7].

Property 1 The VI is a metric. *(1) $VI(\mathcal{C},\mathcal{C}')$ is always non-negative and $VI(\mathcal{C},\mathcal{C}') = 0$ if and only if $\mathcal{C} = \mathcal{C}'$. (2) $VI(\mathcal{C},\mathcal{C}') = VI(\mathcal{C}',\mathcal{C})$ (3) (Triangle inequality) For any 3 clusterings $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ of D*

$$VI(\mathcal{C}_1,\mathcal{C}_2) + VI(\mathcal{C}_2,\mathcal{C}_3) \;\geq\; VI(\mathcal{C}_1,\mathcal{C}_3) \tag{20}$$

The space of all clusterings being finite, the VI 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 a la K-means [6] 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.

Upper bounds. The following 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.*

$$VI(\mathcal{C}, \mathcal{C}') \ \leq \ \log n \tag{21}$$

For example, $\mathcal{C} = \{\{1\}, \{2\}, \{3\}, \dots \{n\}\}$ 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.

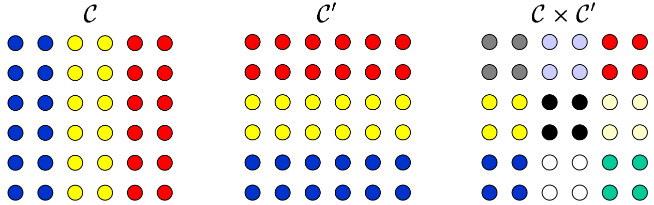
If the number of clusters is bounded by a constant K^* we can derive a bound that is dependent on K^* only.

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 and the same K^* 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 it also makes sense to compare, add, subtract VI distances across the two clustering spaces independently of the sizes of the underlying data sets.

Fig. 2. Two maximally separated clusterings $\mathcal{C}$ and $\mathcal{C}'$, having each $K = 3$ clusters, and their join $\mathcal{C} \times \mathcal{C}'$, having 9 clusters.



The local neighborhood A consequence of having a metric is that we can define ϵ-radius balls around any clustering. The following properties give the distances at which the nearest neighbors of a clustering $\mathcal{C}$ will lie. They also give an intuition of what kind of clusterings lie "immediately near" a given one, or, in other words, 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*

$$P'(k') \;=\; \begin{cases} P(k') & \text{if } C'_{k'} \in \mathcal{C} \\ P(k'|k)P(k) & \text{if } C'_{k'} \;\subseteq\; C_k \;\in\; \mathcal{C} \end{cases} \tag{22}$$

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

$$P(k_l|k) \;=\; \frac{|C'_{k_l}|}{|C_k|} \tag{23}$$

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,

$$VI(\mathcal{C}, \mathcal{C}') = P(k)H_{|k} \tag{24}$$

The same value is obtained when performing the reverse operation, i.e when a set of clusters is merged into a single one. Equation (24) 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.

The next two properties are direct consequences of Property 5.

Property 6 Splitting a cluster into equal parts. *If $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting C_k into q equal clusters, then $VI(\mathcal{C},\mathcal{C}') = P(k)\log q$.*

Property 7 Splitting off one point. *If $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting one point off C_k and making it into a new cluster, then*

$$VI(\mathcal{C},\mathcal{C}') \;=\; \frac{1}{n}[n_k \log n_k - (n_k-1)\log(n_k-1)] \tag{25}$$

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

Property 8 Collinearity of the join. *The triangle inequality holds with equality for two clusterings and their join.*

$$VI(\mathcal{C},\mathcal{C}') \;=\; VI(\mathcal{C},\mathcal{C}\times\mathcal{C}') + VI(\mathcal{C}',\mathcal{C}\times\mathcal{C}') \tag{26}$$

Thus, the join of two clusterings is "collinear" with and "in between" the clusterings in this metric space. 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 9 *For any two clusterings we have*

$$VI(\mathcal{C}, \mathcal{C}') \ \geq \ VI(\mathcal{C}, \mathcal{C} \times \mathcal{C}') \tag{27}$$

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 Properties 7 and 9, 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 $VI(\mathcal{C}, \mathcal{C}') \ \geq \ 2/n$ for $\mathcal{C} \neq \mathcal{C}'$.

The last property implies that the smallest 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, obtaining $n' = \alpha n$ and a new data set D' that has α 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.

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

Property 10 Linearity 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*

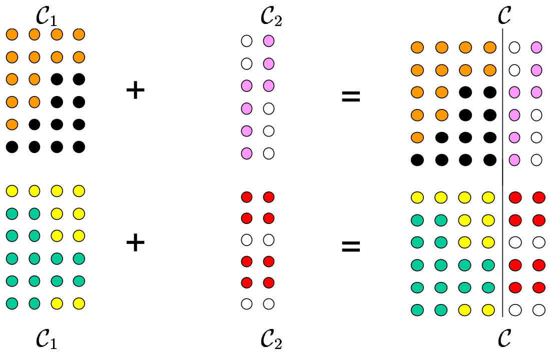
$$VI(\mathcal{C}', \mathcal{C}'') \ = \ \sum_{k=1}^{K} P(k) VI(\mathcal{C}'_k, \mathcal{C}''_k) \tag{28}$$

This property is illustrated in figure 3 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 10 can be seen in another way yet. If two 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 is a linear combination of the VI distances at the level of the component data sets.

Finally, a property pertaining to the computation time of the variation of information.

Fig. 3. Illustration of linearity. If $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$ and $\mathcal{C}' = \mathcal{C}'_1 \cup \mathcal{C}'_2$ then $VI(\mathcal{C}, \mathcal{C}') = \frac{n_1}{n_1+n_2} VI(\mathcal{C}_1, \mathcal{C}'_1) + \frac{n_2}{n_1+n_2} VI(\mathcal{C}_2, \mathcal{C}'_2)$.



Property 11 *$VI(\mathcal{C}, \mathcal{C}')$ can be computed in $\mathcal{O}(n + KK')$ time.*

This is not surprising, since $VI(\mathcal{C}, \mathcal{C}')$, just like the previously presented criteria, is completely determined by the contingency table $[n_{kk'}]$. The first term in the above formula corresponds to the computation of the contingency table, while the second represents the computation of the VI from it.

5 Discussion

5.1 Scaled distances between clusterings

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 give invariance with n particular attention because, in any situation where comparisons are not restricted to a single data set, the value of a criterion that is not n-invariant would be useless without being accompanied by the corresponding n.

The Rand, Fowlkes-Mallows, Jacard, 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. These metrics depend strongly on n but they can be 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 (6) 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, $\mathcal{M}_{inv}$, $\mathcal{D}_{inv}$ for two clusterings with K clusters that are maximally separated under the VI distance. Such a situation is depicted in figure 2. The two clusterings have $n_k = n'_k = n/K$ and $n_{kk'} = n/K^2$ for all $k, k' = 1, \dots K$. It is assumed that n is a multiple of K^2 for simplicity. It can be shown that this pair of clusterings is also maximizing the $\mathcal{D}_{inv}$ and $\mathcal{M}_{inv}$ metrics under the constraint that $K = K'$.

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

$$\begin{aligned} VI^{max} &= 2 \log K \\ \mathcal{D}_{inv}^{max} &= 1 - \frac{1}{K} \\ \mathcal{M}_{inv}^{max} &= \frac{2}{K} - \frac{1}{K^2} \end{aligned} \tag{29}$$

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 $1/K$.

5.2 Linearity and locality

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

Property 12 Linearity 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 (29 showing that the Mirkin metric has to decrease rapidly with the number of clusters. Note also that the unscaled versions of $\mathcal{D}$, $\mathcal{M}$ are additive, hence also linear.

Linearity 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 linear 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

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

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 7.

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 Meilă-Heckerman $\mathcal{H}$ indices are also local. The Larsen, Fowlkes-Mallos and Jacard indices are not local. See [7] for details.

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.

5.3 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 contrast with 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 [4]. 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. If pairwise relationships between data points are fundamental to the current application, then a criterion based on pair counts should be used. Model based clustering and centroid based clustering (e.g the K-means algorithm of [6]) focus not on pairwise relationships but on the relationship between a point and its cluster or centroid. Therefore, the VI distance is a priori better suited with applications of model based clustering than indices based on counting pairs.

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 allow their values to be compared. However, the existing rescaling methods make strong assumptions about the way the clusterings were generated, that are commonly violated in practice. By contrast, the variation of information makes no assumptions about how the clusterings were 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.

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 metrics 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 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.

Last but not least, the variation of information fares well compared to other criteria in that it is easy to understand. For those readers who are familiar with information theory, VI is a natural extension of basic concepts. For the other readers, this paper has given a thorough description of the behavior of VI. The very fact that this paper contains more proved results that any of [1, 3–5, 10–12] and that most results were easy to obtain is an argument for the "understandabilty" and "predictablity" of this metric. In addition to understanding the VI per se, the properties of variation of information presented represent a tool that helps us think about the space of clusterings in a precise way and brings it nearer our intuition.

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.

References

1. 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.

2. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 1991.
3. 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.
4. Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.
5. 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.
6. S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.
7. Marina Meilă. Comparing clusterings. Technical Report 419, University of Washington, 2002. www.stat.washington.edu/reports.
8. Marina Meilă and David Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1/2):9–29, 2001.
9. Boris Mirkin. *Mathematical classification and clustering*. Kluwer Academic Press, 1996.
10. W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.
11. Stijn van Dongen. Performance criteria for graph clustering and Markov cluster experiments. Technical Report INS-R0012, Centrum voor Wiskunde en Informatica, 2000.
12. David L. Wallace. Comment. *Journal of the American Statistical Association*, 78(383):569–576, 1983.