Comparing Clusterings

Marina Meilă
Department of Statistics
University of Washington
mmp@stat.washington.edu

October 20, 2002

Abstract

This paper proposes an information theoretic criterion for comparing two clusterings of the same data set. The criterion, called variation of information (VI), measures the amount of information that is lost or 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 transitive and thus, surprisingly enough, is a true metric on the space of clusterings.

1 Introduction

It is in general a non-trivial task to compare different clusterings of the same data set. Some reasons for this are much the same as for the fruitlessness of postulating a "best" clustering method. Additional reasons have to do with the space of all possible clusterings of a set, which although finite, has a structure complex enough to challenge human intuition.

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 first challenge, 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 clustering task in mind has in these properties a precise description of the criterion's behaviour.

The paper starts with presenting previously used clustering criteria (section 2) and exposing some of the most common criticisms they have received. The variation of information is introduced in section 3 and its properties are presented in section 4. The next section, 5, presents extensions of the criterion to data sets with non-uniform weighting and to soft clusterings. The discussion in section 6 concludes the paper.

2 Related work

In the simplest acception of the term, a clustering $\mathcal{C}$ is a partition of the data set D into sets $C_1, C_2, \ldots C_K$ called *clusters* such that

$$C_k \cap C_l = \emptyset \quad \text{and} \quad \bigcup_{k=1}^{K} C_k = D.$$

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 *con-*

tingency table of the pair $\mathcal{C}, \mathcal{C}'$. The contingency table M 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 $\mathcal{C}'$.

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

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} the number of point pairs that are in different clusters under both $\mathcal{C}$ and $\mathcal{C}'$
- N_{10} the number of point pairs that are in the same cluster under $\mathcal{C}$ but not under $\mathcal{C}'$
- N_{01} the number of point pairs that are 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 M. For example $2N_{11} = \sum_{k,k'} m_{kk'}^2 - n$. See [4] for details.

Wallace [11] proposed the two asymmetric criteria $\mathcal{W}_I, \mathcal{W}_{II}$ below.

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

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

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 $\mathcal{W}_I, \mathcal{W}_{II}$.

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

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 [4]. The index is used by substracting 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. It can be shown that the unnormalized Fowlkes-Mallows index represents a scalar product [2].

The adjusted Rand index is a similar transformation introduced by [5] of Rand's [9] criterion

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

For both indices, the base-line is a function of the number of points in each cluster of the two clusterings. This has two major drawbacks: 1) The baseline needs to be recomputed for every pair of clusterings. This is no more than a computation burden one could put up with, if other more serious problems didn't exist. 2) The value of the baseline for $\mathcal{B}$ 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 (.6 1] [4]. The base-line for the adjusted Rand index varies even more: from 0.5 to 0.95 [4]. 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 $\mathcal{B} = .7$ at a baseline of .4 the same as a value of $\mathcal{B} = .6$ at a baseline of .2? Note that both values, after normalization, yield 0.5 so theoretically they should be equal.

Another problem is that the baseline is an expectation under a null hypothesis. The null hypothesis is that a) the two clusterings are sampled independently b) the clusterings are sampled from the set of all partition pairs with fixed n_k, $n'_{k'}$ points in each cluster [4, 5]. In practice, both assumptions are normally violated. The clusterings are not independent since usually they are obtained by clustering algorithms from the same data set. Many algorithms take a number of clusters K as input but none that I know of take the number of points in each cluster as given. The number 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. In fact, I do not know of any published clustering paper that uses these indices and does not violate one or both assuptions.

The problems exposed above have been known in the statistical community for a long time; see for example [11].

There are other criteria in the literature, to which the above discussion applies. So is the Jacard [2] index

$$\mathcal{J}(\mathcal{C}, \mathcal{C}') = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \quad (6)$$

an improved version of the Rand index, and the Mirkin [8] metric

$$\mathcal{K}(\mathcal{C}, \mathcal{C}') = \sum_k n_k^2 + \sum_{k'} {n'_{k'}}^2 - 2\sum_k \sum_{k'} m_{kk'}^2 \quad (7)$$

The latter is obviously 0 for identical clusterings and positive otherwise. In fact [8], this metric corresponds to the Hamming distance between certain binary vector representations of each partition. This metric can

also be rewritten as

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

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. Larsen [6] uses

$$\mathcal{L}(\mathcal{C},\mathcal{C}') \;=\; \frac{1}{K}\sum_{k}\max_{k'}\frac{2m_{kk'}}{n_k+n'_k} \qquad (8)$$

This is an asymmetric criterion that is 1 when the clusterings are identical.

Meilă and Heckerman [7] 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 $m_{kk'}$ of the contingency table M in decreasing order. The largest of them, be it m_{ab}, entails a match between C_a and C'_b, the second largest not in row a or column b entails the second match, etc. 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

$$\mathcal{H}(\mathcal{C},\mathcal{C}') \;=\; \frac{1}{n}\sum_{k'=match(k)} m_{kk'} \qquad (9)$$

The index is asymmetric and takes value 1 for identical clusterings.

The criteria $\mathcal{L},\mathcal{H}$ present difficulties raised by their assymmetry. 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 stubbornly converges to $\frac{1}{3}$ when $f\to 0$ against our intuition that the difference should be neglijible 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 [10]

$$\mathcal{D}(\mathcal{C},\mathcal{C}') \;=\; 2n-\sum_{k}\max_{k'} m_{kk'}-\sum_{k'}\max_{k} m_{kk'} \qquad (10)$$

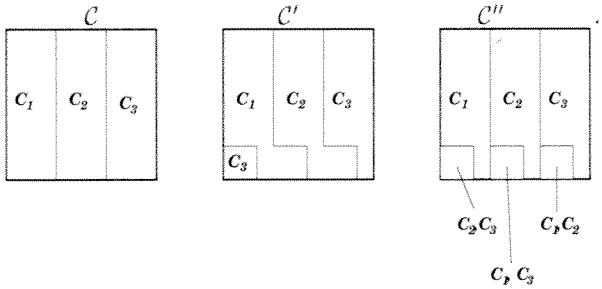

Figure 1: Clustering $\mathcal{C}'$ is obtained from $\mathcal{C}$ by moving a small fraction of the points in each cluster to the next cluster; $\mathcal{C}''$ is obtained from $\mathcal{C}$ by reassigning the same fraction of each cluster equally between all other clusters. Which of $\mathcal{C}'$, $\mathcal{C}''$ is closer to to the original clustering?

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, 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 [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

$$P(k)=\frac{n_k}{n} \qquad (11)$$

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

$$H(\mathcal{C}) \;=\; -\sum_{k=1}^{K} P(k)\log P(k) \qquad (12)$$

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 want to define the *mutual information* between two clusterings, i.e the information that one clustering has about the other. Denote by $P(k)$, $k = 1,\dots K$ and $P'(k')$, $k' = 1,\dots K'$ the random variables associated with the clusterings $\mathcal{C}, \mathcal{C}'$. We introduce yet another distribution, $P(k,k')$ which represents the probability that a point belongs to C_k in clustering $\mathcal{C}$ and to $C'_{k'}$ in $\mathcal{C}'$.

$$P(k,k') \;=\; \frac{|C_k \bigcap C'_{k'}|}{n} \qquad (13)$$

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 random variables induced by the clusterings

$$I(\mathcal{C},\,\mathcal{C}') \;=\; \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k')\log\frac{P(k,k')}{P(k)P'(k')} \qquad (14)$$

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.

$$I(\mathcal{C},\mathcal{C}') \;=\; I(\mathcal{C}',\mathcal{C}) \;\geq\; 0 \qquad (15)$$

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

$$I(\mathcal{C},\mathcal{C}') \;\leq\; \min(H(\mathcal{C}),\; H(\mathcal{C}')) \qquad (16)$$

Equality in the above formula occurs when one clustering completely determines the other. For example, if

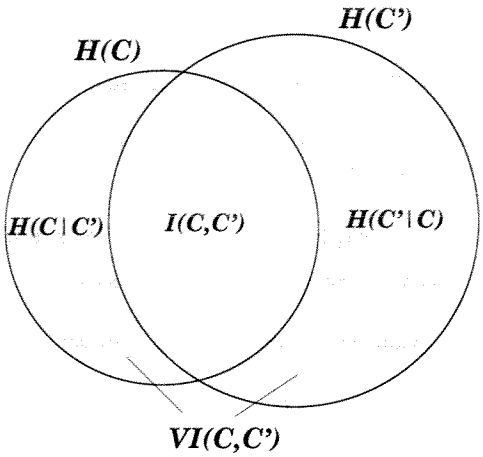

Figure 2: The variation of information and related quantities.

$\mathcal{C}'$ 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

$$VI(\mathcal{C},\mathcal{C}') \;=\; H(\mathcal{C}) + H(\mathcal{C}') - 2I(\mathcal{C},\mathcal{C}') \qquad (17)$$

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

$$VI(\mathcal{C},\mathcal{C}') \;=\; [H(\mathcal{C}) - I(\mathcal{C},\mathcal{C}')] + [H(\mathcal{C}') - I(\mathcal{C},\mathcal{C}')] \qquad (18)$$

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 still have to gain, when going from clustering $\mathcal{C}$ to clustering $\mathcal{C}'$.

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 creates in the space 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 problem, 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.

4.1 The variation of information is a metric

Property 1 Positivity *$VI(\mathcal{C},\mathcal{C}')$ is always positive and $VI(\mathcal{C},\mathcal{C}') = 0$ if and only if $\mathcal{C} = \mathcal{C}'$.*

Property 2 Symmetry. $VI(\mathcal{C},\mathcal{C}') = VI(\mathcal{C}',\mathcal{C})$

Property 3 Triangle inequality. *For any 3 clusterings on D*

$$VI(\mathcal{C}_1,\mathcal{C}_2) + VI(\mathcal{C}_2,\mathcal{C}_3) \geq VI(\mathcal{C}_1,\mathcal{C}_3) \qquad (19)$$

The first three properties imply that VI is a *metric* (or *distance*) on clusterings. Note however that 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 organizing a large set of clusterings. For example, one can design algorithms à la K-means 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 4 *The value of $VI(\mathcal{C},\mathcal{C}')$ depends only on the relative sizes of the clusters. It does not depend on the number of points in the data set.*

Property 5 *The following bound is attained for all n.*

$$VI(\mathcal{C},\mathcal{C}') \leq \log n \qquad (20)$$

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 there are more clusterings available. I.e, 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 6 *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, substract VI distances across the two clustering spaces independently of the sizes of the underlying data sets.

4.3 The local neighborhood

A consequence of having a bounded metric with a known bound is that we can define ϵ-radius balls around any clustering. The upper bound on the metric gives an absolute upper bound on ϵ. Let us now look at the range of small ϵ values. Since the space is discrete, 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 $\mathcal{C}$.

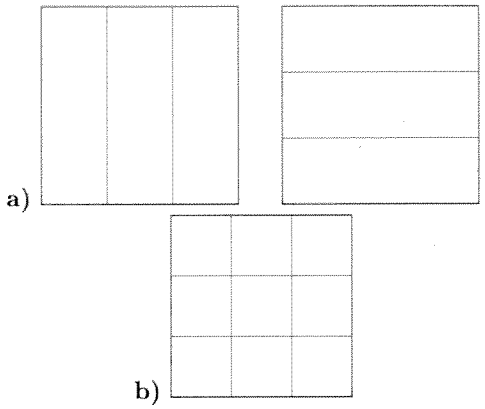

Figure 3: Two maximally separated clusterings, whose $VI(\mathcal{C},\mathcal{C}') = 2\log K$ **(a)** and their product **(b)**.

In addition, it is reasonable to expect that the clusterings near a given $\mathcal{C}$ are obtained by "small" changes to $\mathcal{C}$. Thus, we will also be concerned with the question: what changes to a clustering are small according to the VI distance?

Property 7 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*

$$P'(k') \;=\; \begin{cases} P(k') & \text{if } C'_{k'} \in \mathcal{C} \\ P(k'|k)P(k) & \text{if } C'_{k'} \subseteq C_k \end{cases} \qquad (21)$$

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

$$P(k_l|k) \;=\; \frac{|C'_{k_l}|}{|C_k|} \qquad (22)$$

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,

$$VI(\mathcal{C},\mathcal{C}') = P(k)H_{|k} \qquad (23)$$

The same value is obtained when performing the reverse operation, i.e when a set of clusters is merged into a single one. Equation (23) shows that 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 7.

Property 8 Splitting a cluster into two equal parts *If $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting C_k into two equal clusters, then $VI(\mathcal{C},\mathcal{C}') = P(k)$.*

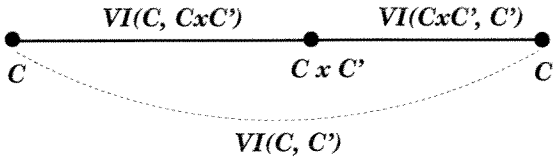

Figure 4: Illustration of Property 10.

Property 9 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*

$$VI(\mathcal{C},\mathcal{C}') \;=\; \frac{1}{n}[n_k\log n_k - (n_k-1)\log(n_k-1)] \qquad (24)$$

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 *product* of clusterings $\mathcal{C}$ and $\mathcal{C}'$ by

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

Hence, the product of two clusterings is the clustering formed from all the nonempty intersections of clusters from $\mathcal{C}$ with clusters from $\mathcal{C}'$. Note that if $\mathcal{C}'$ is a refinement of $\mathcal{C}$, then $\mathcal{C}\times\mathcal{C}' = \mathcal{C}'$.

Property 10 Collinearity of the product *The triangle inequality holds with equality for two clusterings and their product.*

$$VI(\mathcal{C},\mathcal{C}') \;=\; VI(\mathcal{C},\mathcal{C}\times\mathcal{C}') \;+\; VI(\mathcal{C}',\mathcal{C}\times\mathcal{C}') \qquad (25)$$

The proof, elementary, is given in the appendix.

Thus, the product of two clusterings is "collinear" with and "inbetween" 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 11 *For any two clusterings we have*

$$VI(\mathcal{C},\mathcal{C}') \;\geq\; VI(\mathcal{C},\mathcal{C}\times\mathcal{C}') \qquad (26)$$

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 9 and 11, 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

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

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.

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

Property 12 *$VI(\mathcal{C},\mathcal{C}')$ can be computed in $\mathcal{O}(n + KK')$ time.*

This is not surprising, since VI, just like the previously presented criteria, is completely determined by the contingency table M. The first term in the above formula corresponds to the computation of M, while the second represents the computation of the VI from M.

5 Extensions

5.1 Weighted data points

The variation of information distance can be easily extended to the case where each data point is weighted by a non-uniform probability $P(i)$. All one needs to do is to replace the value of $P(k)$, $P'(k)$ and $P(k,k')$ with

$$P(k) = \sum_{i \in C_k} P(i) \quad (27)$$

$$P(k,k') = \sum_{i \in C_k \cap C'_{k'}} P(i) \quad (28)$$

Weighting the data points non-uniformly will distort the distances w.r.t the VI evaluated from a uniform weighting. It does not however change the fundamental properties of the variation of information, which remains a metric. Properties 1, 2, 3, 7, 10 and 11 remain true for non uniform weightings. As for the properties related to upper bounds, a non-uniform $P(i)$ will always have an entropy smaller than $\log n$. Therefore, the upper bound in Property 5 will still hold, but it may not be a tight bound any more. A similar reasoning shows that Property 6 holds too, but may not be tight. For arbitrary non-uniform $P(i)$, there is no lower bound on the distance between two distinct clusterings.

Note that the weighting of the data points needs not be known at the time of the clustering, nor does it have to be consistent with any weighting used to obtain the clusterings.

5.2 Soft clusterings

It is equally easy to define the variation of information distance for soft clusterings. A soft clustering is a clustering where each data point belongs with a certain probability to each of the clusters:

$$Pr[i \in C_k] = \gamma_i(k) \quad (29)$$

Once again, all we need to do is to redefine the joint distribution $P(k,k')$ and the marginals $P(k)$, $P'(k')$. The latter definitions are straightforward

$$P(k) = \sum_{i \in D} P(i)\gamma_i(k) \qquad P'(k') = \sum_{i \in D} P(i)\gamma_i(k') \quad (30)$$

Or, if we assume uniform weighting

$$P(k) = \sum_{i \in D} \gamma_i(k) \qquad P'(k') = \sum_{i \in D} \gamma_i(k') \quad (31)$$

The joint distribution $P(k,k')$ is computed as follows

$$P(k,k') = P(k)P(k'|k) \quad (32)$$

$$= P(k) \sum_{i=1}^{n} P(i|k)P(k'|i) \quad (33)$$

$$= P(k) \sum_{i=1}^{n} \frac{P(i)\gamma_i(k)}{P(k)} \gamma_i(k') \quad (34)$$

$$= \sum_{i=1}^{n} P(i)\gamma_i(k)\gamma_i(k') \quad (35)$$

One can now replace the above results in equations (11) and (13) to obtain the VI distance between two clusterings.

In the above we have made the assumption that

$$P(k'|i,k) = P(k'|i) \quad (36)$$

In other words, we assume that knowing k is irrelevant to knowing k' if the chosen point i is known. If in a certain application this assumption does not hold, then the above computation of $P(k,k')$ should be replaced with one based on the assumptions appropriate to the task on hand.

To decide if assumption (36) is acceptable, one must look at the meaning of the "partial"/"probabilistic"/"soft"/"fuzzy" assignments $\gamma_i(k)$ in the current context. For example, in the case of model based clustering by mixtures of normal distributions (see e.g [1]) the partial assignments $\gamma_i(k)$ are estimated by the EM algorithm. They represent the posterior probability that data point i is generated by the distribution corresponding to cluster k. The implicit assumption is that the choice of k for point i is independent of anything else. Hence, in this case, if we compare clusterings obtained by independent runs of the EM algorithm on the same data set, assumption (36) is justified.

6 Discussion

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 $VI(\mathcal{C},\mathcal{C}'') = VI(\mathcal{C},\mathcal{C}') + 2K\log(K-1)$ implying that $\mathcal{C}'$ is closer to $\mathcal{C}$ than $\mathcal{C}''$ for all $K > 2$.

In contrast to the comparison criteria based on pair count, the variation of information is not directly concerned with relationships between pairs of points, or with triples like [5]. 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 counting pairs is intrinsic to the current task, then a criterion based on pair counts should be used. Otherwise, the variation of information offers advantages that are not matched by the counting pairs criteria.

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 exisiting 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, in a sense that has been discussed before, 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.

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, in contrast to the previous one, will preserve comparability of distances across data sets, independently of the number of data points in each set. Therefore, we recommend it as a further simplification of the criterion.

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 Jacard 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. 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 bring it nearer our intuition.

Proofs

Proof of Property 3 The *conditional entropy* between two clusterings $\mathcal{C}, \mathcal{C}'$ is defined as

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

Thus, $VI(\mathcal{C}, \mathcal{C}') = H(\mathcal{C}|\mathcal{C}') + H(\mathcal{C}'|\mathcal{C})$.

We prove that for any clusterings $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$

$$H_{2|1} + H_{3|2} - H_{3|1} \;\geq\; 0 \tag{37}$$

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

$$\begin{aligned} H_{2|1} + H_{3|2} - H_{3|1} &\geq H_{2|1} + H_{3|2,1} - H_{3|1} && (38)\\ &= H_{2,3|1} - H_{3|1} && (39)\\ &\geq 0 \end{aligned}$$

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

$$H_{2|1} + H_{1|2} + H_{3|2} + H_{2|3} - H_{3|1} - H_{1|3} \;\geq\; 0 \tag{40}$$

which is the triangle inequality for the variation of information.

Proof of Property 10

First we prove that

$$H(\mathcal{C}|\mathcal{C}') \;=\; H(\mathcal{C} \times \mathcal{C}'|\mathcal{C}') \tag{41}$$

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

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

Then,

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

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

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

References

[1] Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, September 1993.

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

[6] 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.

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

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

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

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

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