Comparing Clusterings – An Axiomatic View

Marina Meilă MMP@STAT.WASHINGTON.EDU
University of Washington, Box 354322, Seattle, WA 98195-4322

Abstract

This paper views clusterings as elements of a lattice. Distances between clusterings are analyzed in their relationship to the lattice. From this vantage point, we first give an axiomatic characterization of some criteria for comparing clusterings, including the variation of information and the unadjusted Rand index. Then we study other distances between partitions w.r.t these axioms and prove an impossibility result: there is no "sensible" criterion for comparing clusterings that is simultaneously (1) aligned with the lattice of partitions, (2) convexely additive, and (3) bounded.

1. Introduction

This paper views clusterings as elements of a lattice, the natural algebraic structure for the partitions of a set. A criterion for comparing clusterings is thus seen as a function of pairs of elements in the lattice. The goal in doing this is to contribute to the better understanding of the space of all clusterings of a set and of the problem of comparing clusterings.

This task is is an important component in the evaluation of clustering algorithms. A clustering, or a clustering algorithm, can be evaluated by *internal criteria*, e.g distortion, likelihood, that are generally problem and algorithm dependent. There is another kind of evaluation, called *external* evaluation, that simply measures how close is the obtained clustering to a gold standard clustering. The comparison criteria studied here are all external. As such, they are independent of the algorithm or of the way the clusterings were obtained.

There are many competing criteria for comparing clusterings, with no clear best choice. In fact, as this paper will underscore, it is probably meaningless to search for a best criterion for comparing clusterings, just as it is meaningless to search for the best clustering algorithm. Algorithms are "good" in as much as they match the task at hand. With respect to distances between clusterings, our goal is to better understand their properties, their limitations, and the implied assumptions underlying them. This paper does so via an axiomatic study of several distances.

We first discuss some desirable properties of distances between clusterings that amount to decomposing them additively over elementary operations on clusterings like splitting a cluster, merging two data sets, etc. Then we give an axiomatic characterization for one criterion, the variation of information. The choice is not incidental: this criterion is closely matched to the lattice of partitions, as it will be shown. It will turn out that the resulting axioms have each an intuitive interpretation.

The axiomatic framework introduced is extended to characterize other distances between clusterings (the Mirkin metric, the Rand index and the van Dongen metric). We also discuss the "classification error" clustering distance. Finally, we derive an impossibility result for comparing clusterings: we show that no distance on the space of partitions can simultaneously satisfy three desirable properties, each of which makes the distance intuitive in some sense.

2. Additive properties for distances and the lattice of partitions

This section introduces the basic notation, then introduces the lattice of partitions and and lists some properties related to the lattice representation which will be relevant in the rest of the paper.

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

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

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

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 $n_k \geq 1$; 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 $n'_{k'}$. The two clusterings may have different numbers of clusters. To compare $\mathcal{C}$ and $\mathcal{C}'$ is to define a symmetric, non-negative function $d(\mathcal{C}, \mathcal{C}')$ that measures how different are the two clusterings. If $\mathcal{C} = \mathcal{C}'$ then $d(\mathcal{C}, \mathcal{C}') = 0$. We shall call the function d a *distance* although in general it may not satisfy the triangle inequality. If the triangle inequality is satisfied, then d is called a *metric*.

It should also be noted that this definition is not aligned to the majority of clustering comparison criteria, like the Rand (Rand, 1971), Fowlkes-Mallows (Fowlkes & Mallows, 1983) and other *indices*. These are taking values in [0, 1], with larger values as the clusterings become more similar (and being 1 when $\mathcal{C} = \mathcal{C}'$). This will not preclude us to consider these indices, as one can turn any index $i(\mathcal{C}, \mathcal{C}')$ into a distance by simply setting $d(\mathcal{C}, \mathcal{C}') = 1 - i(\mathcal{C}, \mathcal{C}')$.

The number of points in the intersection of clusters C_k of $\mathcal{C}$ and $C'_{k'}$ of $\mathcal{C}'$ is denoted $n_{kk'}$.

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

A distance d between that depends only on the relative values $n_{kk'}/n$ and does not directly depend on n is said to be *n-invariant*.

The lattice of partitions One can represent all clusterings of a finite set D as the nodes of a graph, as illustrated by figure 1; 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 a dataset D forms a lattice called the *lattice of partitions* (Stanley, 1997). The graph just described is known as the *Hasse diagram*[1] of this lattice. At the top of the diagram is the clustering with just one cluster, denoted by $\hat{1} = \{D\}$. At the bottom is $\hat{0} = \{\{1\}, \{2\}, \ldots \{n\}\}$ the clustering with n clusters each containing a single point. For all but the smallest n, the space of all clusterings, although finite, is huge (superexponential in n (Stanley, 1997)); a graphical representation like the Hasse diagram can aid the understanding of the complex relationships between and thus guide us in choosing or designing the most relevant distances on this space. In particular, we may ask if there are distances between clusterings that are "aligned" with the lattice, in the sense that $d(\mathcal{C}, \mathcal{C}')$ can be expressed as a sum of distances along edges of the lattice?

In the following we give a formal definition of three *additivity* properties, all related to the lattice of partitions, and we explain why they are desirable.

Additivity w.r.t refinement (AR) If $\mathcal{C}'$ is obtained from $\mathcal{C}$ by splitting one or more clusters, then we say that $\mathcal{C}'$ is a refinement of $\mathcal{C}$. A distance d is *additive w.r.t refinement* iff for any clusterings $\mathcal{C}, \mathcal{C}', \mathcal{C}''$ such that $\mathcal{C}'$ is a refinement of $\mathcal{C}$ and $\mathcal{C}''$ a refinement of $\mathcal{C}'$ we have

$$d(\mathcal{C}, \mathcal{C}'') = d(\mathcal{C}, \mathcal{C}') + d(\mathcal{C}', \mathcal{C}'') \tag{1}$$

For instance, $\mathcal{C}'' = \{\{a\}, \{b\}, \{c\}, \{d\}\}$ is a refinement of $\mathcal{C}' = \{\{a\}, \{b\}, \{c, d\}\}$, which in turn is a refinement of $\mathcal{C} = \{\{a, b\}, \{c, d\}\}$. The AR property says that the distance $d(\mathcal{C}, \mathcal{C}'')$ is a sum of the distances corresponding to the two succesive refinements that transform $\mathcal{C}$ into $\mathcal{C}''$. Cluster splitting corresponds to taking *downward* steps in the Hasse diagram.

Additivity w.r.t the join (AJ). The *join* of clusterings $\mathcal{C}$ and $\mathcal{C}'$ is defined as

$$\mathcal{C} \times \mathcal{C}' = \{C_k \cap C'_{k'} \,|\, C_k \in \mathcal{C},\, C'_{k'} \in \mathcal{C}',\, C_k \cap C'_{k'} \neq \emptyset\} \tag{2}$$

Hence, the join of two clusterings is the clustering formed from all the nonempty intersections of clusters from $\mathcal{C}$ with clusters from $\mathcal{C}'$. A distance d is *additive w.r.t to the join* iff for any clusterings $\mathcal{C}, \mathcal{C}'$

$$d(\mathcal{C}, \mathcal{C}') = d(\mathcal{C}, \mathcal{C} \times \mathcal{C}') + d(\mathcal{C}', \mathcal{C} \times \mathcal{C}') \tag{3}$$

This property is relevant for clusterings $\mathcal{C}, \mathcal{C}'$ which are not a refinement of each other. One can think of obtaining such a $\mathcal{C}'$ from $\mathcal{C}$ by a series of cluster splits (downward steps along the lattice edges) followed by a set of cluster mergers (upward steps in the lattice). A distance d is AJ if it can be expressed as a sum of distances along such a path. Note that usually there are several possible paths between two clusterings (see for instance $\{\{a\}, \{b, c, d\}\}$ and $\{\{a, c\}, \{b\}, \{d\}\}$ in figure 1); the sum is the same no matter what path is taken.

From a practical point of view, AJ also means that for any two clusterings $\mathcal{C}, \mathcal{C}'$ there is a clustering $\mathcal{C} \times \mathcal{C}$ (and usually others) which has more clusters than both $\mathcal{C}$ and $\mathcal{C}'$ but is closer to both of them than $d(\mathcal{C}, \mathcal{C}')$. In a geometric sense, the join $\mathcal{C}, \mathcal{C}'$ is "on the line segment" between $\mathcal{C}, \mathcal{C}'$. If in an application two clusterings should be close only if the number of clusters K, K' are nearly equal, then the distance d should *not* be AJ. However, there are other

[1] The emphasized terms in this section represent standard lattice terminology. Their precise definitions can be found in e.g (Stanley, 1997).

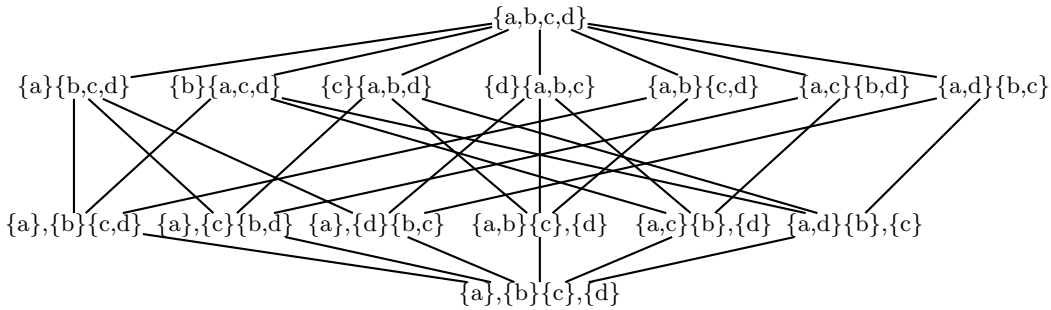

Figure 1. The lattice of partitions of D ={a,b,c,d}. Note that $\hat{1}$ ={{a,b,c,d}}, $\hat{0}$ ={{a},{b},{c},{d}}; the clusterings $\hat{1}$, {{a},{b, c, d}}, {{a},{b},{c, d}}, $\hat{0}$ are collinear according to A3; the clusterings {{a},{b, c, d}}, {{a},{b},{c, d}}, {{a,b},{c, d}} are collinear according to A4; the clusterings {{a},{b, c, d}}, {{a},{b},{c, d}}, {{b},{a, c, d}} and $\hat{1}$ are on a closed straight line; and there are 3 straight lines from {{d},{a, b, c}} to $\hat{0}$.

applications, image segmentation being a prime example, where breaking off a set b of pixels from a segment $\{a, b\}$ is less of a mistake than breaking them off and attaching them to another segment c (i.e we want $d(\{\{a, b\}, \{c\}\}, \{\{a\}, \{b\}, \{c\}\}) < d(\{\{a, b\}, \{c\}\}, \{\{a\}, \{b, c\}\})$). In such applications, the number of segments is less meaningful than the grouping of the pixels; therefore AJ is reasonable as it represents the total error as as sum of two kinds of error.

Convex additivity (CA). 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 $P(k)$ represent the proportion of data points that belong to cluster C_k. Then

$$d(\mathcal{C}', \mathcal{C}'') \;=\; \sum_{k=1}^{K} P(k) d(\mathcal{C}'_k, \mathcal{C}''_k) \qquad (4)$$

This property expresses additivity of d over the sublattices corresponding to the individual clusters C_k. In particular, CA implies that if only some cluster(s) of $\mathcal{C}$ are changed to obtain $\mathcal{C}'$, then then $d(\mathcal{C}, \mathcal{C}')$ depends only on the affected clusters, and is independent on how the unaffected part of D is partitioned. For an example, consider the pairs $\mathcal{C} = \{\{a, b\}, \{c, d\}\}$, $\mathcal{C}' = \mathcal{C} = \{\{a\}, \{b\}, \{c, d\}\}$ and $\tilde{\mathcal{C}} = \{\{a, b\}, \{c\}, \{d\}\}$, $\tilde{\mathcal{C}} = \{\{a\}, \{b\}, \{c\}, \{d\}\}$. In both cases the first cluster is split while the remaining cluster(s) are left unchanged. Therefore, if d is CA then $d(\mathcal{C}, \mathcal{C}') = d(\tilde{\mathcal{C}}, \tilde{\mathcal{C}}')$.

Properties AJ and CA were described in (Meilă, 2003), where several distances were discussed w.r.t satisfying them. In this section we extend and deepen the geometric view suggested there. The remainder of the paper addresses the reverse question: instead of checking whether a given distance satisfies a certain property, we will ask what properties uniquely define a certain distance. The focus will be on the additive properties discussed above. Hence the complementary question: how can one characterize all the distances that satisfy them?

The reason we pay special attention to the properties AV, CA and AJ in this paper, is because splitting a cluster, forming a clustering by taking the union of two clusterings on two subsets of the data, and merging two clusters are (sequences of) elementary and intuitive operations that one can do on clusterings. It is easy and natural to think of changing a clustering $\mathcal{C}$ into $\mathcal{C}'$ by appying these operations one after the other. If the distance d satisfies AV, CA and AJ, then d will measure the change between $\mathcal{C}$ and $\mathcal{C}'$ as a sum of elementary changes, each corresponding to one step.

Finally, if d is also a metric, some additional geometric insights are possible. One can extend the notion of a *straight line* from Euclidean space to a metric space: 3 points in a metric space for which triangle inequality is satisfied with equality are said to lie on a straight line, in other words to be *collinear*. Additivity w.r.t refinement implies that the clusterings along a vertical *chain* of lattice edges are collinear. As all "vertical" straight lines meet both at $\hat{1}$ and $\hat{0}$, this space is clearly non-Euclidean. In general, 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 1 illustrates these properties on a simple example. An even more interesting picture is implied by the "horizontal" straight lines that exist because of the additivity w.r.t the joint. These lines are composed of the vertical "descending" segment $\mathcal{C}, \mathcal{C} \times \mathcal{C}'$, continued by

the "ascending" segment $\mathcal{C} \times \mathcal{C}', \mathcal{C}'$. Figure 1 shows an example of such a straight line. Moreover, using both properties, one can derive that the union of any two chains that have the same endpoints forms a "closed" straight line according to such a metric.

3. The axioms of the variation of information

The *variation of information* (d_{VI}), introduced in (Meilă, 2003), measures the distance between two clusterings in terms of the information difference between them.

$$d_{VI}(\mathcal{C}, \mathcal{C}') \;=\; H(\mathcal{C}) + H(\mathcal{C}') - 2I(\mathcal{C}, \mathcal{C}') \qquad (5)$$

where H and I represent respectively the entropies[2] of and the mutual information[3] between the two clusterings.

In (Meilă, 2003) the variation of information was shown to satisfy AJ and AC and that it is a metric, among several other properties. It is also easy to show that it satisfies additivity w.r.t refinement (see the appendix).

Now we show that a subset of these properties uniquely defines d_{VI}.

Theorem 1 *The variation of information is the unique cluster comparison criterion d that satisfies the axioms:*

A1 **Symmetry** *For any two clusterings $\mathcal{C}, \mathcal{C}'$*

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

A2 **Additivity w.r.t refinement** *Denote by $\hat{0}$ and $\hat{1}$ the unique clusterings having $K = n$ respectively $K = 1$ clusters. For any clustering $\mathcal{C}$*

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

A3 **Additivity w.r.t the join** *For any two clusterings $\mathcal{C}, \mathcal{C}'$*

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

A4 **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*

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

A5 **Scale** *Denote by $\mathcal{C}_K^U$ the "uniform" clustering, i.e the clustering with K equal clusters. If $\mathcal{C}_K^U$ exists, then*

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

In the above, $\hat{1}_{n_k}$ is the $\hat{1}$ clustering of the dataset C_k containing n_k points. The proof of the theorem is constructive and is given in the appendix. Note that the axioms do not require d to be a metric; this follows implicitly.

Intuitively, axioms A2 and A3 describe the geometric properties of d_{VI}, i.e that it is aligned with the lattice of partitions. Axiom A2 is a weak version of the AR property defined in the previous section. Axioms A4 and A5 set the scale of d and in particular its logarithmic growth rate. They are reminiscent of the postulates III and IV of entropy as given in (Rènyi, 1970).

It is interesting to see what happens if the last two 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. This is what we are going to do in the next section.

4. Other metrics for comparing clusterings

The clustering literature contains quite a number of criteria for comparing clusterings: the Rand index (Rand, 1971), the Jaccard index (Ben-Hur et al., 2002), the Folwkes-Mallows index (Fowlkes & Mallows, 1983), the Huber and Arabie indices (Hubert & Arabie, 1985), the Mirkin metric (Mirkin, 1996), the Van Dongen metric (van Dongen, 2000), as well as statistically "adjusted" versions of some of the above (Hubert & Arabie, 1985).

We shall start with two criteria, the Mirkin and Van Dongen metrics, for which we give an axiomatic characterization.

The Mirkin metric is defined by (Mirkin, 1996)

$$d'_M(\mathcal{C}, \mathcal{C}') \;=\; \sum_k n_k^2 + \sum_{k'} {n'_{k'}}^2 - 2\sum_k \sum_{k'} n_{kk'}^2 \qquad (6)$$

This metric can also be rewritten (Ben-Hur et al., 2002) as

$$d'_M(\mathcal{C}, \mathcal{C}') \;=\; 2N_{disagree}(\mathcal{C}, \mathcal{C}') \qquad (7)$$

where $N_{disagree}$ is defined as the number of point pairs which are in the same cluster under $\mathcal{C}$ but in different clusters under $\mathcal{C}'$ or viceversa. The Rand index is de-

[2] $H(\mathcal{C}) \;=\; -\sum_{k=1}^{K} \frac{n_k}{n} \log \frac{n_k}{n}$

[3] $I(\mathcal{C}, \mathcal{C}') \;=\; \sum_{k=1}^{K} \sum_{k'=1}^{K'} \frac{n_{k,k'}}{n} \log \frac{n_{k,k'}}{n} \frac{n_k}{n} \frac{n'_{k'}}{n}$

fined as

$$i_R(\mathcal{C},\mathcal{C}') \;=\; \frac{n(n-1)-2N_{disagree}(\mathcal{C},\mathcal{C}')}{n(n-1)} \qquad (8)$$

Therefore the characterization of d_M below will reflect immediately on the (unadjusted) Rand index. In what follows we shall use an rescaled form of the Mirkin metric which is n-invariant and bounded.

$$d_M(\mathcal{C},\mathcal{C}') \;=\; \frac{d'_M(\mathcal{C},\mathcal{C}')}{n^2} \qquad (9)$$

The Van Dongen criterion was also proved to be a metric (van Dongen, 2000)

$$d'_D(\mathcal{C},\mathcal{C}') \;=\; 2n - \sum_k \max_{k'} n_{kk'} - \sum_{k'} \max_k n_{kk'} \quad (10)$$

As with the Mirkin metric, we shall use its bounded n-invariant version

$$d_D(\mathcal{C},\mathcal{C}') \;=\; \frac{d'_D(\mathcal{C},\mathcal{C}')}{2n} \qquad (11)$$

To proceed with our axiomatic study, we first compute the distances between $\hat{1}$ and $\mathcal{C}_K^U$ under the invariant Van Dongen and Mirkin metrics.

$$d_D(\hat{1},\mathcal{C}_K^U) \;=\; \frac{1}{2}\left(1-\frac{1}{K}\right)$$
$$d_M(\hat{1},\mathcal{C}_K^U) \;=\; 1-\frac{1}{K}$$

With respect to convex additivity, it is easy to prove (see also (Meilă, 2003)) that d_D is convexely additive, while the Mirkin metric satisfies

$$d_M(\mathcal{C},\mathcal{C}') \;=\; \sum_{k=1}^{K} \frac{n_k^2}{n^2} d_M(\hat{1}_{n_k},\mathcal{C}'_k) \qquad (12)$$

whenever $\mathcal{C}'$ is a refinement of $\mathcal{C}$ and $\mathcal{C}'_k$ represents the partitioning induced by $\mathcal{C}'$ on cluster C_k of $\mathcal{C}$.

We now can replace the scale axioms of d_{VI} with axioms corresponding to d_D and d_M respectively. We obtain a negative

Theorem 2 *There is no cluster comparison criterion that satisfies axioms A1–A4 and*
A5.D $d(\hat{1},\mathcal{C}_K^U) \;=\; 1-1/K$

Theorem 3 *The unique cluster comparison criterion d that satisfies axioms A1, A3, A4, and*

$$A2.D \qquad d(\hat{1},\mathcal{C}) \;=\; \frac{1}{2}\left[1-\frac{\max_k n_k}{n}\right]$$

is the invariant van Dongen metric d_D.

Theorem 4 *The unique cluster comparison criterion d that satisfies axioms A1–A3, A5.D and*

A4.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 d_M.

Thus, the invariant Mirkin metric is also aligned with the lattice of partitions. The additivity axiom A4.M has several consequences. First, it shows that d_M is local, i.e changes inside one cluster depend only on the relative size of that cluster and of the nature of the change, and are not affected by how the rest of the data set is partitioned. But the union of two or more datasets shrinks distances (because of the weighting with the squares of the proportions), a rather counter-intuitive behavior. It is worth recalling that all these properties of the Mirkin metric are readily translated into similar properties of the unadjusted Rand index. The "unnatural" behavior of the Rand index with increasing K has been noted early on and is the main reason why this index is used mostly in its adjusted form (Hubert & Arabie, 1985).

The van Dongen metric is horizontally aligned with the lattice of partitions but not vertically, that is it satisfies axioms A1, A3, A4, A5.D. To uniquely characterize it, we introduced axiom A2.D, which implies A5.D but is significantly stronger[4].

How about other, more popular indices for comparing partitions? The space does not permit us to describe and compare all of them here (such comparisons can be found in (Ben-Hur et al., 2002; Hubert & Arabie, 1985; Meilă, 2003; Wallace, 1983) and others). One fact shown in (Meilă, 2003) is that the Jaccard, Fowlkes-Mallows, some of the Huber and Arabie and all the adjusted indices, including the widely used adjusted Rand index, are non-local (that is, a change inside a single cluster counts differently depending on how the rest of the data is clustered). Therefore they will rate worse on the scale of understandability and in particular cannot satisfy the convex additivity property. Also, it is easy to show that most of the above indices are only asymptotically n-invariant, so their axiomatic description would be much more complicated (and consequently a much less illuminating exercise).

[4] We have not proved that A2.D is the *weakest possible* axiom that uniquely deterimines d_D. But we believe this as likely and will consider it in further research.

Another very interesting criterion, the classification error, is discussed in the next section.

5. The classification error metric

The classification error (CE) distance d_{CE} is defined as

$$d_{CE}(\mathcal{C},\mathcal{C}') \;=\; 1-\frac{1}{n}\max_{\sigma}\sum_{k=1}^{K} n_{k,\sigma(k)} \qquad (13)$$

In the above, it is assumed w.l.o.g that $K \leq K'$, σ 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 σ 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 d_{CE}.

From the computational point of view, it is not necessary to explicitly enumerate all correspondences (order $K!$). The maximum can be computed in polinomial time as the solution of a linear program identical to the maximum bipartite matching algorithm in graph theory (Golumbic, 1980). The CE distance is simple and intuitive, especially if the two clusterings are close together. The following lemma describes its properties from the point of view of the axioms considered in this paper.

Lemma 5 *The classification error distance d_{CE} satisfies axioms A1 (symmetry), A4 (convex additivity), A5.D (scales like $1-1/K$). It violates axioms A2 and A3, hence it is not aligned with the lattice of partitions.*

The proof is sufficiently simple and has been omitted. The above result shows similarities betweed d_{CE} and especially the d_D metric, and underscores its convex additivity, locality, and n-invariance, all of which contribute to making d_{CE} a most intutitive criterion. The dissimilarity is of course the non-alignment with the lattice of partitions. We have not yet found a complete characterization of d_{CE}, this is a matter of further research.

6. An impossibility result for comparing partitions

Theorem 2 prompts the question: what kind of scalings in A5 are compatible with A1–A4? To answer this question, we change A5 to the weaker A5.H:

A5.H $d(\hat{1},\mathcal{C}_K^U) \;=\; h(K)$ where h is a non-decreasing function of K.

Then, the result below shows that A5 is essentially superfluous.

Theorem 6 *Any clustering comparison criterion satisfying A1–A4 and A5.H is identical to d_{VI} up to a multiplicative constant.*

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

From theorem 6 the following impossibility result follows immediately.

Corrollary 7 *There is no d symmetric, n-invariant, with $d(\hat{1},\mathcal{C}_K^U)$ non-decreasing, that satisfies simultaneously the following three properties:*
- *d is aligned to the lattice of partitions (axioms A2, A3)*
- *d is convexly additive (axiom A4)*
- *d is bounded*

Hence, there is no criterion for comparing clusterings that can satisfy all three of the above desirable properties. The users of distances between clusterings will have to make choices. What are the tradeoffs?

The first property gives geometric intuition in addition to the intuition one would get from having a metric. Preserving this property would be useful in designing search algorithms in the space of clusterings and proving their properties.

The second one is, in our opinion, the most important for the "understandability" of a distance, because it is a law of composition. It says that under unions of sets, the distances on the parts are weighted in proportion to the sizes of the parts and then summed.

The argument for the third property is partly "historical". The vast majority of criteria for comparing clusterings are bounded between 0 and 1. Moreover, in statistics, the tradition is to indicate identity between two clusterings by a 1 (like in the Rand, Fowlkes-Mallows, Jaccard indices and their adjusted versions). Another very appealing reason to use a distance that is bounded (for instance between 0 and 1) is to interpret it as a probability. When can we do this? The aforementioned indices can all be interpreted as probabilities (see the original papers (Rand, 1971; Fowlkes & Mallows, 1983; Hubert & Arabie, 1985; Ben-Hur et al., 2002) for details), but their adjusted versions can not (Hubert & Arabie, 1985). In this respect, perhaps the most interpretable is the classification error d_{CE}, which is indeed the optimal error probability if clustering was regarded as classification.

Note however that a criterion that is bounded between 0 and 1 carries the implicit assumption that clusterings can only get negligibly more diverse as the number of clusters increases. Thus, while using d_{CE} may be the most natural and intuitive when K is small, this distance will loose its resolution power for large K. Whether a bounded or unbounded criterion for comparing clusterings is better depends ultimately on the clustering application at hand. This paper's aim in this respect is to underscore the possible choices and their consequences.

7. Conclusion

This is the first axiomatic approach to comparing clusterings. What are the benfits of such an exercise? Characterizing distances between clusterings in terms of axioms highlights the essential properties of these distances, from which all others follow. For example, we have seen that being aligned to the lattice of partitions is a very strong requirement: together with an additivity constraint (like A4 or A4.M), it completely determines the values of a distance on the lattice.

An impossibility result is also important, because it reveals an essential characteristic of the problem itself. In this case, very loosely speaking, we have shown that the lattice of partitions, which is the space where all possible clusterings lie, has too rich a structure to be packed inside a bounded ball without breaking some of the structure. For an analogy, one can say that the Euclidean n-dimensional space, another space with rich structure, cannot be folded into a sphere without losing some of its properties. Thus, presenting distances between clusterings in an axiomatic framework helps better understand their properties and their limitations. It is interesting to note that an impossibility result for clustering exists in (Kleinberg, 2002). The Kleinberg paper refers to *clustering criteria* (e.g single linkage, mean-squared error, etc) and shows that there is none that satisfies three desirable properties.

We have given a prominent role to the lattice of partitions, interpreting distances as functions on pairs of points in the lattice. We believe this view will also become useful in the future, whether one will use distances aligned to the lattice like the d_{VI}, or completely unaligned like d_{CE}. Clustering is a hard problem and seeing clusterings as nodes in a graph opens the possibility of applying new techniques, based on graph and lattice theory, to this task.

Acknowledgements

Thanks to Nilesh Dalvi and Pavel Krivitsky who each gave a proof for lemma 8 and to the anonymous reviewer who pointed me to the work of C. Rajski. This research was partly supported by NSF grant IIS-0313339.

Proofs

Proof of AR for the d_{VI} distance. d_{VI} can be expressed as a sum of conditional entropies (Meilă, 2003)

$$d_{VI}(\mathcal{C},\mathcal{C}') = H(\mathcal{C}|\mathcal{C}') + H(\mathcal{C}'|\mathcal{C}') \tag{14}$$

The proof then follows from elementary properties of the conditional entropy (see (Cover & Thomas, 1991) for details): first, if $\mathcal{C}'$ is a refinement of $\mathcal{C}$, then $H(\mathcal{C}|\mathcal{C}') = 0$; then, one applies the chain rule for conditional entropy.

Proof of Theorem 1 From A4 and A5 we have that

$$\begin{aligned} d(\hat{0},\mathcal{C}) &= \sum_k \frac{n_k}{n} d(\hat{1},\mathcal{C}^U_{n_k}) && (15)\\ &= \sum_k \frac{n_k}{n} \log n_k && (16)\\ &= \sum_k \frac{n_k}{n} (\log \frac{n_k}{n} + \log n) && (17)\\ &= \log n - H(\mathcal{C}) && (18) \end{aligned}$$

From A2 we get $d(\hat{1},\mathcal{C}) = \log n - d(\hat{0},\mathcal{C}) = H(\mathcal{C})$. For any two clusterings $\mathcal{C},\mathcal{C}'$ define by $\mathcal{C}_k$ the clustering induced by $\mathcal{C}'$ on $C_k \in \mathcal{C}$.

$$\begin{aligned} d(\mathcal{C},\mathcal{C}\times\mathcal{C}') &= \sum_{k=1}^{K} \frac{n_k}{n} d(\hat{1},\mathcal{C}_k) && (19)\\ &= \sum_{k=1}^{K} \frac{n_k}{n} H(\mathcal{C}_k) && (20)\\ &= H(\mathcal{C}|\mathcal{C}') && (21) \end{aligned}$$

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

Proof of Theorem 2

$$\begin{aligned} d(\hat{0},\mathcal{C}) &= \sum_k \frac{n_k}{n} d(\hat{1},\mathcal{C}^U_{n_k}) && (22)\\ &= \sum_k \frac{n_k}{n} \left(1 - \frac{1}{n_k}\right) && (23)\\ &= 1 - \frac{K}{n} && (24) \end{aligned}$$

Therefore $d(\hat{1},\mathcal{C}) = (1-1/n)-(1-K/n) = (K-1)/n$ if $|\mathcal{C}| = K$. This contradicts A5 according to which $d(\hat{1},\mathcal{C}^U_K) = (K-1)/K$.

Proof of Theorems 3 and 4 These proofs follow the same steps as the proof of Theorem 1 and are therefore omitted.

Proof of Theorem 6 We have consecutively:

$$\begin{aligned} d(\hat{1},\hat{0}) &= h(n) \quad \text{by A5.H} && (25)\\ d(\hat{1},\mathcal{C}) &= h(n) - d(\hat{0},\mathcal{C}) \quad \text{by A2} && (26)\\ d(\hat{0},\mathcal{C}) &= \sum_k \frac{n_k}{n} h(n_k) \quad \text{by A4} && (27)\\ d(\hat{1},\mathcal{C}_K^U) &= h(n) - d(\hat{0},\mathcal{C}_K^U) && (28)\\ &= h(n) - K\frac{1}{K}h(\frac{n}{K}) && (29) \end{aligned}$$

Since $n/K = M$ is an integer, and recalling A5.H we can rewrite the last equality as

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

or equivalently

$$h(KM) \;=\; h(K) + h(M) \qquad (30)$$

for any positive integers K, M. By lemma 8 below, this implies that $h(n) = C \log n$ for all $n = 1, 2, 3, \ldots$.

It follows that A1-A4 together with A5.H imply essentially the original A5 (up to the multiplicative constant C) and therefore d cannot be but proportional to the VI.

Lemma 8 *Let $h : \{1, 2, \ldots\} \to [0, \infty)$ be a non-decreasing function satisfying (30) for any positive integers K, M. Then $h(n) = C \log n$ for any n.*

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

$$\begin{aligned} h(2^{\lfloor a \rfloor}) &\le h(2^a) = h(n^k) \le h(2^{\lceil a \rceil}) && (31)\\ \lfloor a \rfloor C &\le qh(n); \le \lceil a \rceil C && (32)\\ \frac{\lfloor a \rfloor}{a} &\le \tfrac{h(n)}{C \log n}; \le \frac{\lceil a \rceil}{a} && (33) \end{aligned}$$

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

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing* (pp. 6–17).

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* Wiley.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*, 553–569.

Golumbic, M. (1980). *Algorithmic graph theory and perfect graphs.* Academic Press, New York.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Kleinberg, J. (2002). An impossibility theorem for clustering. *Advances in Neural Information Processing Systems.* Cambridge, MA: MIT Press.

Meilă, M. (2003). Comparing clusterings by the variation of information. *Proceedings of the Sixteenth Annual Conference ofn Computational Learning Theory (COLT).* Springer.

Mirkin, B. (1996). *Mathematical classification and clustering.* Kluwer Academic Press.

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

Rènyi, A. (1970). *Probability theory.* North-Holland.

Stanley, R. P. (1997). *Enumerative combinatorics.* Cambridge Unversity Press.

van Dongen, S. (2000). *Performance criteria for graph clustering and Markov cluster experiments* (Technical Report INS-R0012). Centrum voor Wiskunde en Informatica.

Wallace, D. L. (1983). Comment. *Journal of the American Statistical Association*, *78*, 569–576.