Monotonic Entropies - Distribution Measures for Partitions of Data

Prof. Dan A. Simovici
University of Massachusetts Boston

1. A Historic Perspective

2. Set Partitions

3. Axiomatization of Monotonic Partition Entropy

4. Conditional Monotonic Entropy

5. An Application: External Validation of Clustering

The notion of entropy, the cornerstone of information theory, was introduced by Claude Shannon in his 1948 double paper in Bell System Technical Journal, as a limit of on lossless data compression in a noiseless data transmission channel.

- There exists an ample literature containing axiomatizations of the notion of entropy for probability distributions.
- Some of these axiomatizations involve the Shannon entropy (Khinchin, Rényi, Fadeev,). Others, focus on generalizations of entropy (Daroczy, Havrda, Tsallis, Furuichi, Simovici).

We present on an axiomatization of entropy that leverages algebraic properties of sets of partitions of finite sets in order to produce a simpler system of axioms for entropy, and to extend this notion to a diverse collection of data types.

Partitions are fundamental for clustering algorithms were we seek to detect groupings of objects that have similar properties or are geometrically close to each other.

There is a vast literature that focuses on clustering algorithms and a great diversity of approaches to clustering.

Evaluating cluster quality is an important and challenging task for comparing appropriateness of clustering algorithms for various object configurations.

Why partitions?

Partitions of sets are more expressive than probability distributions.

Definition

Let S be a non-empty set. A partition of S is a collection of non-empty subsets $\{B_i \mid i \in I\}$ of S such that for $i, j \in I$ we have $B_i \cap B_j = \emptyset$, and $\bigcup_{i \in I} B_i = S$.

The set of partitions on S is denoted by $PART(S)$.

A probability distribution that describe a 3-valued random variable:

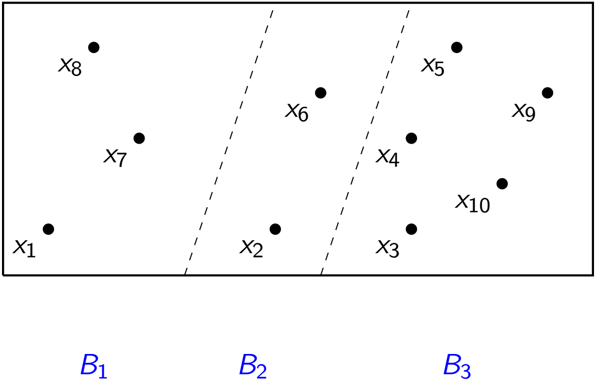
$$p : (0.3 \;\; 0.2 \; 0.5)$$

A three-block partition:

$$\pi = \{\{x_1, x_7, x_8\}, \{x_2, x_6\}, \{x_3, x_4, x_5, x_9, x_{10}\}\}$$

Obviously, we can extract the probability distribution from the partition; the reverse process will not work because there are many partitions that correspond to a probability distribution.

Partition $\pi \in PART(\{x_1, \ldots, x_9\})$:



Compare the probability distribution $(3/10, 2/10, 5/10)$ with the partition π shown in the previous picture:

- blocks are $B_1 = \{x_1, x_7, x_8\}$, $B_2 = \{x_2, x_6\}$, $B_3 = \{x_3, x_4, x_5, x_9, x_{10}\}$;
- if the elements belong to a metric space, various metric parameters (centroids, diameters, etc.) could be considered;
- if x_i are vertices of a graph and various edges exist between these points, the configuration of these edges may help establish properties of partitions.

Partitions have a natural partial order and the set of partitions has a rich algebraic structure.
If $\pi \in PART(S)$ and $x, y \in S$ belong to the same block of π we write $x \equiv y(\pi)$. The relation "$\equiv$" is reflexive, symmetric, and transitive and, therefore, it is an equivalence relation on S. Conversely, if ρ is an equivalence of S, the sets of the form $[x]_\rho = \{u \in S \mid (x, u) \in \rho\}$ constitute a partition π_ρ of S.

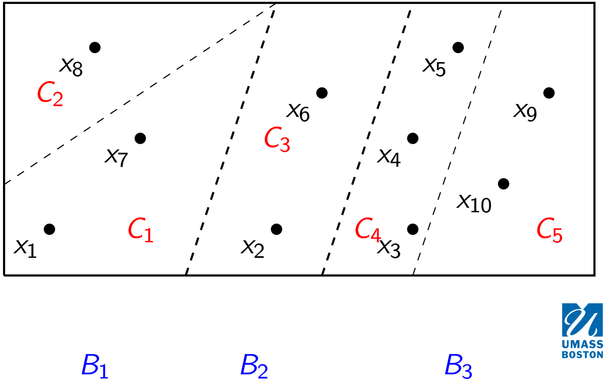
A partial order "$\leqslant$" is defined on partitions in $PART(S)$ by setting $\pi \leqslant \sigma$ if each block of π is included in a block of σ.

Example

Let σ be the partitions whose blocks are:

$$C_1 = \{x_1, x_7\}, C_2 = \{x_8\}, C_3 = \{x_2, x_6\}, C_4 = \{x_3, x_4, x_5\}, C_5 = \{x_9, x_{10}\}$$

We have $C_1 \subseteq B_1$, $C_2 \subseteq B_1$, $C_3 = B_2$, $C_4, C_5 \subseteq B_3$, hence $\pi \subseteq \sigma$.

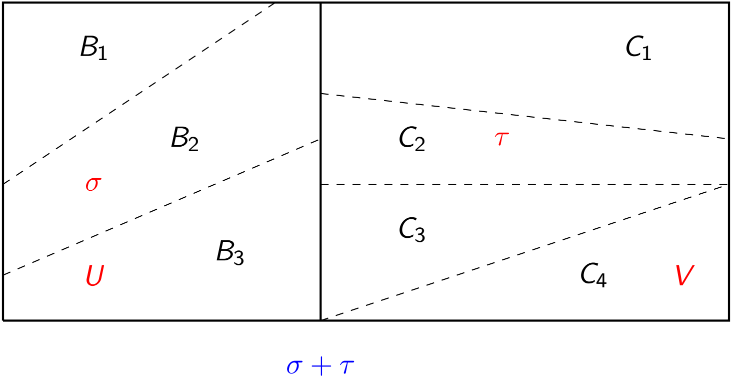


- The partition $\alpha_S = \{\{x\} \mid x \in S\}$ is the least element of the partially ordered set $(PART(S), \leqslant)$.
- The one-block partition $\omega_S = \{S\}$ is the largest element of $(PART(S), \leqslant)$.

If $\pi, \sigma \in PART(S)$, $\pi \leqslant \sigma$, and there is no partition $\tau \in PART(S) - \{\pi, \sigma\}$ such that $\pi \leqslant \tau \leqslant \sigma$, then we say that σ *covers* π and we write $\pi \lessdot \sigma$. It is easy to show that $\pi \lessdot \sigma$ if and only if σ is obtained from π by fusing two of the blocks of π.

Let U, V be two non-empty, disjoint sets, and let $\sigma \in PART(U)$, and $\tau \in PART(V)$, where $\sigma = \{B_1, \ldots, B_m\}$ and $\tau = \{C_1, \ldots, C_n\}$. The *sum of the partitions* σ *and* τ is the partition $\sigma + \tau \in PART(U \cup V)$ defined as:

$$\sigma + \tau = \{B_1, \ldots, B_m, C_1, \ldots, C_n\}.$$



For every two non-empty disjoint sets U and V we have:

$$\begin{aligned} \alpha_U + \alpha_V &= \alpha_{U \cup V}, \\ \omega_U + \omega_V &= \{U, V\} \in PART(U \cup V). \end{aligned}$$

Furthermore, if U, V, W are non-empty disjoint sets, $\sigma \in PART(U)$, $\tau \in PART(V)$ and $\upsilon \in PART(W)$, we have

$$\sigma + (\tau + \upsilon) = (\sigma + \tau) + \upsilon,$$

a property referred to as the *restricted associativity* of partition addition. The term "restricted" refers to the fact that the underlying sets U, V, W are supposed to be disjoint.

If $\sigma = \{B_1, \ldots, B_m\} \in PART(S)$ then we have:

$$\sigma = \omega_{B_1} + \cdots + \omega_{B_m}.$$

If the set S consists of a single element, $S = \{s\}$, then $\alpha_S = \omega_S = \{s\}$.

Our axiomatization of partition entropies starts with monotonic functions defined on sets of partitions. We present three examples of monotonic functions defined on specialized collections of sets that will allow us to generate a variety of entropy types.

Let $\mu : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ be a *non-negative monotonic* function of sets, that is, a function such that $U \subseteq V$ implies $\mu(U) \leqslant \mu(V)$ for $U, V \in \mathcal{P}(S)$, and $|U| > 1$ implies $\mu(U) > 0$.

Example

Let S be a finite set and let $\mu : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ be given by $\mu(B) = |B|^{\beta}$ for $B \in \mathcal{P}(S)$ and some $\beta > 0$. The function is clearly monotonic and $B \neq \emptyset$ implies $\mu(B) > 0$.
Furthermore, if $|B| = 1$, then $\mu(B) = 1$.

Example

Let $W = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subseteq \mathbb{R}^n$ be a finite set and let d be a metric on $\mathbb{R}^n$.
Define the *centroid* of W as $\mathbf{c}_W = \frac{1}{|W|} \sum_{\mathbf{x} \in W} \mathbf{x}$.
The *sum of square errors* of the set W is defined as:

$$sse(W) = \sum_{i=1}^{m} d^2(\mathbf{x}_i, \mathbf{c}_W) = \sum_{\mathbf{x} \in W} \parallel \mathbf{x} \parallel^2 - |W| \parallel \mathbf{c}_W \parallel^2 .$$

Example cont'd

Example

If W is a finite subset of $\mathbb{R}^n$ and $\sigma = \{U, V\}$ is a bipartition of W a straightforward computation yields:

$$sse(W) = sse(U) + sse(V) + \frac{|U|\,|V|}{|W|} \parallel \mathbf{c}_U - \mathbf{c}_V \parallel^2 .$$

This implies

$$sse(U) + sse(V) \leqslant sse(W). \tag{1}$$

Note also that U, W are two subsets of $\mathbb{R}^n$ such that $U \subseteq W$, we have $sse(U) \leqslant sse(W)$, which shows that *sse* is a monotonic function. Furthermore, if $|W| = 1$, then $\mu(W) = 0$.

Example cont'd

Example

Another function that can be defined on finite subsets of $(\mathbb{R}^n, d)$ is the diameter $\mathrm{diam} : \mathcal{P}(\mathbb{R}^n) \longrightarrow \mathbb{R}_{\geqslant 0}$, given by $\mathrm{diam}(W) = \max\{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x}, \mathbf{y} \in W\}$. It is immediate that diam is monotonic.

Example

Let $G = (V, E)$ be a connected loop-free finite graph having V as its set of vertices and E as its set of edges. For a set of vertices B define $\mathrm{int}(B)$, the *set of internal edges of* B as

$$\mathrm{int}(B) = \{\{x, y\} \in E \mid \{x, y\} \subseteq B\}.$$

This definition is extended to partitions of sets of vertices by defining

$$\mathrm{int}(\pi) = \bigcup_{B \in \pi} \mathrm{int}(B).$$

The set $\mathrm{int}(\pi)$ is the *set of internal edges* of π.

Example

If $\pi, \sigma \in PART(V)$ then $\mathrm{int}(\pi \wedge \sigma) = \mathrm{int}(\pi) \cap \mathrm{int}(\sigma)$.
The set $\mathrm{ext}(\pi)$ of *external edges* of π (also known as *cut edges* of π) consists of edges that join vertices in distinct blocks and is given by:

$$\mathrm{ext}(\pi) = E - \mathrm{int}(\pi).$$

Thus, we have:

$$\begin{aligned}\mathrm{ext}(\pi \wedge \sigma) &= E - \mathrm{int}(\pi \wedge \sigma) \\ &= E - (\mathrm{int}(\pi) \cap \mathrm{int}(\sigma)) \\ &= (E - \mathrm{int}(\pi)) \cup (E - \mathrm{int}(\sigma)) \\ &= \mathrm{ext}(\pi) \cup \mathrm{ext}(\sigma).\end{aligned}$$

Example

Note that

$$\begin{array}{rclcrcl} \mathrm{int}(\alpha_V) & = & \emptyset, & \mathrm{ext}(\alpha_V) & = & E, \\ \mathrm{int}(\omega_V) & = & E, & \mathrm{ext}(\omega_V) & = & \emptyset \end{array}$$

for every graph $G = (V, E)$.
It follows from the above discussion that the function

$$\mathrm{int} : PART(V) \longrightarrow \mathcal{P}(E)$$

is monotonic, while

$$\mathrm{ext} : PART(V) \longrightarrow \mathcal{P}(E)$$

is dually monotonic.

Starting from monotonic functions of sets we introduce a set of three axioms that define an entropy associated to these functions.

Definition

Let S be a non-empty set and let $\mu : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ be a non-negative monotonic function defined on the subsets of S. A *μ-entropy* is a function $\mathcal{H}_\mu : PART(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ that satisfies the following conditions:

- ($\mathbf{A}_0$)-initialization axiom: For any set S, $\mathcal{H}_\mu(\omega_S) = 0$.
- ($\mathbf{A}_1$)-monotonicity axiom: If $\pi, \sigma \in PART(S)$ and $\pi \leqslant \sigma$, then $\mathcal{H}_\mu(\pi) \geqslant \mathcal{H}_\mu(\sigma)$.
- ($\mathbf{A}_2$)-addition axiom: For every finite disjoint subsets U, V of a set S such that $S = U \cup V$, $\sigma \in PART(U)$ and $\tau \in PART(V)$ we have:

$$\mathcal{H}_\mu(\sigma + \tau) = \frac{\mu(U)}{\mu(U \cup V)}\mathcal{H}_\mu(\sigma) + \frac{\mu(V)}{\mu(U \cup V)}\mathcal{H}_\mu(\tau) + \mathcal{H}_\mu(\{U, V\}).$$

Lemma

If $|S| = 1$, *then* $\mathcal{H}_\mu(\alpha_S) = 0$.

Proof.

This follows from the fact that for a singleton set $S = \{a\}$ we have $\alpha_S = \omega_S$. □

UMASS BOSTON

Lemma

Let U, V be two non-empty, finite disjoint sets, $\mu : \mathcal{P}(U \cup V) \longrightarrow \mathbb{R}_{\geqslant 0}$ be a positive monotonic function of sets, and let σ be a partition of the set U. Then,

$$\mathcal{H}_\mu(\sigma + \alpha_V) = \mathcal{H}_\mu(\sigma + \omega_V) + \frac{\mu(V)}{\mu(U \cup V)}\mathcal{H}_\mu(\alpha_V).$$

Proof.

By Lemma 11 we can write:

$$\begin{aligned}
\mathcal{H}_\mu(\sigma + \alpha_V) &= \frac{\mu(U)}{\mu(U \cup V)}\mathcal{H}_\mu(\sigma) + \frac{\mu(V)}{\mu(U \cup V)}\mathcal{H}_\mu(\alpha_V) + \mathcal{H}_\mu(\{S, T\}), \\
\mathcal{H}_\mu(\sigma + \omega_V) &= \frac{\mu(U)}{\mu(U \cup V)}\mathcal{H}_\mu(\sigma) + \mathcal{H}_\mu(\{U, V\}).
\end{aligned}$$

The equalities imply the desired result. □

Theorem

Let S be set such that $|S| \geqslant 2$ and let $\pi = \{B_1, \ldots, B_m\}$ be a partition of S. For any non-negative monotonic function $\mu : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ we have:

$$\mathcal{H}_\mu(\pi) = \mathcal{H}_\mu(\alpha_S) - \sum_{i=1}^{m} \frac{\mu(B_i)}{\mu(S)} \mathcal{H}_\mu(\alpha_{B_i}). \tag{2}$$

Proof

Since $\pi = \omega_{B_1} + \omega_{B_2} + \cdots + \omega_{B_n}$, we can consider the descending sequence of partitions of the set S:

$$\begin{array}{rcl}
\pi_0 & = & \omega_{B_1} + \omega_{B_2} + \cdots + \omega_{B_m} = \pi \\
\pi_1 & = & \alpha_{B_1} + \omega_{B_2} + \cdots + \omega_{B_m} \\
\pi_2 & = & \alpha_{B_1} + \alpha_{B_2} + \cdots + \omega_{B_m} \\
 & \vdots & \\
\pi_m & = & \alpha_{B_1} + \alpha_{B_2} + \cdots + \alpha_{B_m} = \alpha_S.
\end{array}$$

Define $\sigma_i = \alpha_{B_1} + \cdots + \alpha_{B_i} + \omega_{B_{i+2}} + \cdots + \omega_{B_m} \in PART(S - B_{i+1})$ for $1 \leqslant i \leqslant m - 1$. Note that

$$\pi_i = \sigma_i + \omega_{B_{i+1}} \text{ and } \pi_{i+1} = \sigma_i + \alpha_{B_{i+1}}$$

are both partitions of the set S.

Proof cont'd

By a previous Lemma we have:

$$\begin{aligned}
\mathcal{H}_\mu(\pi_1) &= \mathcal{H}_\mu(\pi_0) + \frac{\mu(B_1)}{\mu(S)}\mathcal{H}_\mu(\alpha_{B_1}), \\
\mathcal{H}_\mu(\pi_2) &= \mathcal{H}_\mu(\pi_1) + \frac{\mu(B_2)}{\mu(S)}\mathcal{H}_\mu(\alpha_{B_2}), \\
&\vdots \\
\mathcal{H}_\mu(\pi_m) &= \mathcal{H}_\mu(\pi_{m-1}) + \frac{\mu(B_m)}{\mu(S)}\mathcal{H}_\mu(\alpha_{B_m}).
\end{aligned}$$

Therefore,

$$\mathcal{H}_\mu(\pi_m) = \mathcal{H}_\mu(\pi_0) + \sum_{i=1}^{m} \frac{\mu(B_i)}{\mu(S)}\mathcal{H}_\mu(\alpha_{B_i})$$

Equivalently, since $\pi_m = \alpha_S$, we gave

$$\mathcal{H}_\mu(\pi) = \mathcal{H}_\mu(\alpha_S) - \sum_{i=1}^{m} \frac{\mu(B_i)}{\mu(S)}\mathcal{H}_\mu(\alpha_{B_i}).$$

Corollary

Let S be set such that $|S| \geqslant 2$. For any non-negative monotonic function $\mu : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ and any partition $\pi = \{B_1, \ldots, B_n\} \in PART(S)$ we have:

$$\mathcal{H}_\mu(\alpha_S) \geqslant \sum_{i=1}^{m} \frac{\mu(B_i)}{\mu(S)} \mathcal{H}_\mu(\alpha_{B_i}). \tag{3}$$

Proof.

By the initialization and monotonicity axioms $\pi \leqslant \omega_S$ imply $\mathcal{H}_\mu(\pi) \geqslant \mathcal{H}_\mu(\omega_S) = 0$, hence the μ-entropy of any partition is non-negative. This fact combined with Theorem 13 yields the desired result. □

Example

Let $\mu(S) = |S|^\beta$ for any finite and non-empty set S and $\beta > 0$ and let

$$\mathcal{H}_\mu(\alpha_S) = \frac{1 - |S|^{1-\beta}}{1 - 2^{1-\beta}}.$$

By Theorem 13 this choice of $\mathcal{H}_\mu(\alpha_S)$ implies:

$$\mathcal{H}_\mu(\pi) = \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{B \in \pi} \frac{|B|^\beta}{|S|^\beta}\right),$$

which is the Havrda-Charvàt generalized entropy.

Note that

$$\lim_{\beta \longrightarrow 1+} \frac{1 - |S|^{1-\beta}}{1 - 2^{1-\beta}} = \ln |S|,$$

by a straightforward application of l'Hospital rule.
If $\pi \leqslant \sigma$ the axiom $(\mathbf{A}_1)$ is satisfied. It suffices to show that $\pi \lhd \sigma$ implies $\mathcal{H}_\mu(\pi) \geqslant \mathcal{H}_\mu(\sigma)$, so let $\pi = \{B_1, \ldots, B_{m-2}, B_{m-1}, B_m\}$ and let $\sigma = \{B_1, \ldots, B_{m-2}, B_{m-1} \cup B_m\}$. These choices imply:

$$\begin{aligned}
\mathcal{H}_\mu(\pi) &= \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{i=1}^{m} \frac{|B_i|^\beta}{|S|^\beta}\right), \\
\mathcal{H}_\mu(\sigma) &= \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{i=1}^{m-2} \frac{|B_i|^\beta}{|S|^\beta}\right) - \frac{|B_{m-1} \cup B_m|^\beta}{|S|^\beta},
\end{aligned}$$

and the axiom $(\mathbf{A}_1)$ is satisfied because

$$|B_{m-1}|^\beta + |B_m|^\beta \leqslant |B_{m-1} \cup B_m|^\beta.$$

The special case $\beta = 2$ yields

$$\mathcal{H}_\mu(\pi) = 2\left(1 - \sum_{i=1}^{m} \frac{|B_i|^2}{|S|^2}\right),$$

which is the double of the Gini index.
By applying l'Hospital rule we obtain:

$$\lim_{\beta \to 1} \mathcal{H}_\mu(\pi) = -\sum_{i=1}^{m} \frac{|B_i|}{|S|} \ln \frac{|B_i|}{|S|}$$

which is the Shannon entropy.

Example

Let μ be the positive monotonic function, $\mu(B) = sse(B)$, where B is a finite subset of $\mathbb{R}^n$. Choose $\mathcal{H}_\mu(\alpha_U) = 1$ for every finite set $U \in \mathcal{P}(S)$.
The μ-entropy is:

$$\mathcal{H}_\mu(\pi) = 1 - \sum_{i=1}^{m} \frac{sse(B_i)}{sse(S)},$$

which is the expression of the inertial entropy of a partition.
The satisfaction of axiom ($\mathbf{A}_1$) follows from the inequality $sse(U) + sse(V) \leqslant sse(U \cup V)$.
With the alternative choice, $\mu(B) = \text{diam}(B)$ we obtain the entropy

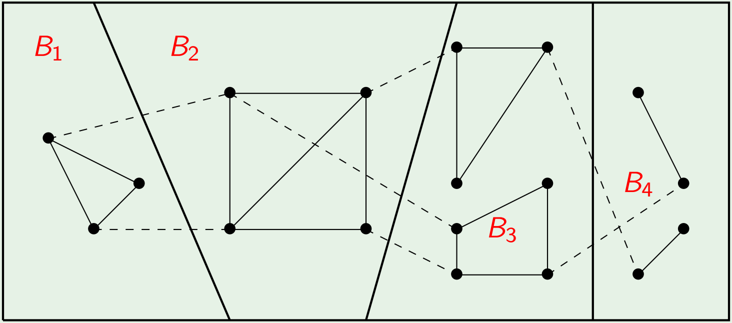
$$\mathcal{H}_\mu(\pi) = 1 - \sum_{i=1}^{m} \frac{\text{diam}(B_i)}{\text{diam}(S)},$$

Example

If $G = (V, E)$ is a connected loop-free finite graph having V as its set of vertices and E as its set of edges, and $\mu(B) = |\text{int}(B)|$ for every set of vertices B, then, choosing $\mathcal{H}_\mu(\alpha_B) = 1$ the expression of μ-entropy of a partition $\pi \in PART(E)$ is:

$$\mathcal{H}_\mu(\pi) = 1 - \sum_{i=1}^{m} \frac{|\text{int}(B_i)|}{|\text{int}(V)|} = \frac{|\text{ext}(\pi)|}{|E|}.$$

Example



$$\mathcal{H}_\mu(\pi) = \frac{|\mathrm{ext}(\pi)|}{|E|} = \frac{7}{26}.$$

Elementary properties of partition cut-sets of graphs allow us to obtain the necessity of axiom $\mathbf{A}_2$ for graph entropies. Indeed, let $\kappa = \{U, W\}$ be a cut in the graph G and let $\sigma \in PART(U)$ and $\tau \in PART(W)$ be two partitions of the sets U and W. The partition $\sigma + \tau$ of V consists of all blocks of σ and all blocks of τ.

An external edge e of partition $\sigma + \tau$ may fall in one of the following pairwise disjoint sets:

- e is an external edge of σ but an internal edge of κ;
- e is an external edge of τ but an internal edge of κ;
- e is an external edge of κ.

Since the sets $\mathrm{ext}(\sigma)$, $\mathrm{ext}(\tau)$, and $\mathrm{ext}(\kappa)$ are disjoint we have:

$$\mathrm{ext}(\sigma+\tau)=\mathrm{ext}(\sigma)\cup\mathrm{ext}(\tau)\cup\mathrm{ext}(\kappa).$$

The last equality implies

$$\frac{|\mathrm{ext}(\sigma+\tau)|}{|V|}=\frac{|U|}{|V|}\frac{|\mathrm{ext}(\sigma)|}{|U|}+\frac{|W|}{|V|}\frac{|\mathrm{ext}(\tau)|}{|W|}+\frac{|\mathrm{ext}(\kappa)|}{|V|}.$$

When this equality is expressed using the graph entropy we recover axiom $\mathbf{A}_2$, namely:

$$\mathcal{H}_\mu(\sigma+\tau)=\frac{\mu(U)}{\mu(U\cup W)}\mathcal{H}_\mu(\sigma)+\frac{\mu(W)}{\mu(U\cup W)}\mathcal{H}_\mu(\tau)+\mathcal{H}_\mu(\{U,W\}).$$

A graph $G = (V, E)$ is bipartite if there exists a bipartition $\pi = \{V_1, V_2\}$ such that $\text{ext}(\pi) = E$. This is equivalent to the existence of a bipartition π such that $\mathcal{H}_\mu(\pi) = 1$.

In general, a graph $G = (V, E)$ is k-colorable, if it has a partition $\pi = \{B_1, \ldots, B_k\}$ such that if $\{x, y\} \in E$, then x and y belong to two distinct blocks of π. In other words, G is k-colorable if and only if there exists a partition of V having k blocks such that $\mathcal{H}_\mu(\pi) = 1$.

Since the graph k-coloring problem is known to be NP-complete, it follows by direct transformation, that the problem of the existence of a partition with k blocks of the set of vertices of a graph and has monotonic entropy equal to 1 is NP-complete.

Let $\pi = \{B_1, \ldots, B_m\} \in PART(S)$ and let $C \subseteq S$. The *trace* of π on C is the partition

$$\pi_C = \{B \cap C \mid B \in \pi \text{ and } B \cap C \neq \emptyset\} \in PART(C).$$

Definition

Let $\pi, \sigma \in PART(S)$, where $\sigma = \{C_1, \ldots, C_n\}$. The μ-conditional entropy of π and σ is given by:

$$\mathcal{H}_\mu(\pi|\sigma) = \sum_{j=1}^{n} \frac{\mu(C_j)}{\mu(S)} \mathcal{H}_\mu(\pi_{C_j}).$$

Note that $\mathcal{H}(\pi|\omega_S) = \mathcal{H}(\pi)$,

$$\mathcal{H}(\omega_S|\sigma) = \sum_{j=1}^{n} \frac{\mu(C_j)}{\mu(S)} \mathcal{H}_\mu(C_j),$$

and $\mathcal{H}_\mu(\pi|\alpha_S) = 0$ for every $\pi \in PART(S)$.

Theorem

For conditional entropy of two partitions $\pi, \sigma \in PART(S)$ we have

$$\mathcal{H}_\mu(\pi|\sigma) = \mathcal{H}_\mu(\pi \wedge \sigma) - \mathcal{H}_\mu(\sigma).$$

Proof

For $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$ in $PART(S)$ the conditional entropy can be written as:

$$\begin{aligned}
\mathcal{H}_\mu(\pi|\sigma) &= \sum_{j=1}^{n} \frac{\mu(C_j)}{\mu(S)} \mathcal{H}_\mu(\pi_{C_j}) \\
&= \sum_{j=1}^{n} \frac{\mu(C_j)}{\mu(S)} \left(\mathcal{H}_\mu(\alpha_{C_j}) - \sum_{i=1}^{m} \frac{\mu(B_i \cap C_j)}{\mu(C_j)} \mathcal{H}_\mu(\alpha_{B_i \cap C_j}) \right) \\
&= \sum_{j=1}^{n} \frac{\mu(C_j)}{\mu(S)} \mathcal{H}_\mu(\alpha_{C_j}) - \sum_{i=1}^{m} \frac{\mu(B_i \cap C_j)}{\mu(S)} \mathcal{H}_\mu(\alpha_{B_i \cap C_j}) \\
&= \mathcal{H}_\mu(\pi \wedge \sigma) - \mathcal{H}_\mu(\sigma).
\end{aligned}$$

UMASS BOSTON

Theorem

The mapping $\delta_\mu : PART(S)^2 \longrightarrow \mathbb{R}_{\geqslant 0}$ *defined as*

$$\delta_\mu(\pi, \sigma) = \frac{d_\mu(\pi, \sigma)}{\mathcal{H}_\mu(\pi \wedge \sigma)}$$

is a metric on $PART(S)$ *such that* $0 \leqslant \delta_\mu(\pi, \sigma) \leqslant 1$ *for* $\pi, \sigma \in PART(S)$.

Proof

The non-negativity and the symmetry of δ_μ are immediate. To prove the triangular axiom we write:

$$\begin{aligned}
\delta_\mu(\pi,\tau) &= \frac{d_\mu(\pi,\tau)}{\mathcal{H}_\mu(\pi\wedge\tau)} \\
&= \frac{\mathcal{H}_\mu(\pi|\tau)+\mathcal{H}_\mu(\tau|\pi)}{\mathcal{H}_\mu(\pi\wedge\tau)} \\
&\leqslant \frac{\mathcal{H}_\mu(\pi|\sigma)}{\mathcal{H}_\mu(\pi\wedge\sigma)}+\frac{\mathcal{H}_\mu(\sigma|\tau)}{\mathcal{H}_\mu(\sigma\wedge\tau)} \\
&\quad +\frac{\mathcal{H}_\mu(\tau|\sigma)}{\mathcal{H}_\mu(\tau\wedge\sigma)}+\frac{\mathcal{H}_\mu(\sigma|\pi)}{\mathcal{H}_\mu(\sigma\wedge\pi)} \\
&= \delta_\mu(\pi,\sigma)+\delta_\mu(\sigma,\pi),
\end{aligned}$$

Furthermore, since $\mathcal{H}_\mu(\pi|\sigma)\leqslant\mathcal{H}_\mu(\pi\wedge\sigma)$ it follows that $0\leqslant\delta_\mu(\pi,\sigma)\leqslant 1$.

For $\pi, \sigma \in \mathit{PART}(S)$ we have:

$$\delta_\mu(\pi, \sigma) = 2 - \frac{\mathcal{H}_\mu(\pi) + \mathcal{H}_\mu(\sigma)}{\mathcal{H}_\mu(\pi \wedge \sigma)}.$$

Example

For the graph-related entropy the distance δ_μ is given by:

$$\delta_\mu(\pi, \sigma) = 2 - \frac{|\mathsf{ext}(\pi)| + |\mathsf{ext}(\sigma)|}{|\mathsf{ext}(\pi \wedge \sigma)|}.$$

We consider an analog of the Rand distance between partitions of sets of edges in undirected graphs that can be introduced using our approach. Let $\pi, \sigma \in \mathit{PART}(V)$; the *graph Rand distance* $\delta(\pi, \sigma)$ between these partitions is:

$$\begin{aligned}\delta(\pi, \sigma) &= 2 - \frac{|\mathsf{ext}(\pi)| + |\mathsf{ext}(\sigma)|}{|\mathsf{ext}(\pi \wedge \sigma)|} \\ &= \frac{|\mathsf{int}(\pi) \cap \mathsf{ext}(\sigma)| + |\mathsf{int}(\sigma) \cap \mathsf{ext}(\pi)|}{|\mathsf{ext}(\pi \wedge \sigma)|},\end{aligned}$$

because

$$\begin{aligned}\mathsf{ext}(\pi \wedge \sigma) - \mathsf{ext}(\pi) &= \mathsf{int}(\pi) \wedge \mathsf{ext}(\sigma) \\ \mathsf{ext}(\pi \wedge \sigma) - \mathsf{ext}(\sigma) &= \mathsf{int}(\sigma) \wedge \mathsf{ext}(\pi).\end{aligned}$$

External clustering validation involves clustering data that is labeled. Labels denote the correct cluster where a data item belongs and this define a *ground truth partition* σ.

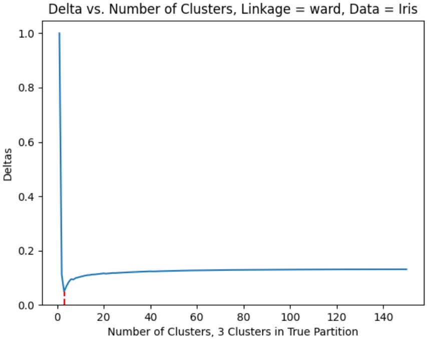
- A clustering algorithm produces a partition π, and the goal of this type of validation is to determine to what extend the partition π is consistent with the ground truth partition σ.
- The consistency is evaluated by using the normalized distance $\delta(\pi, \sigma) \in [0, 1]$. The smaller values of δ (close to 0) mean that the clustering algorithm produces a result consistent with the ground truth partition.
- This approach allows us to determine the correct number of clusters by determining the minimum of $\delta(\pi, \sigma)$ when the parameters that define the partition π are variable.

We examine the effectiveness of the partition distance δ as a method of partition validation. We initially do so on the *iris* data set of the UC Irvine Machine Learning Repository which has three natural classes that form the reference partition σ.

The partition π is obtained via a hierarchical clustering, and the quality of various hierarchical clusterings is evaluated through the distance $\delta(\pi, \sigma)$.

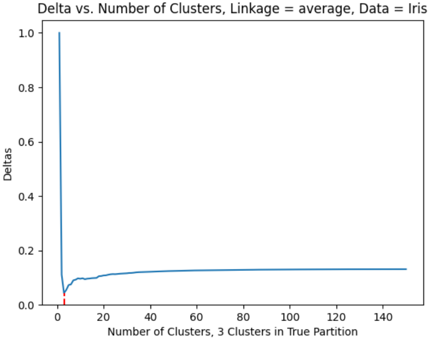
For each partition π we plot its values against the number of clusters in that partition.

For the well-known iris data set who has three clusters we obtain the following results using a variety of hierarchical clustering algorithms:

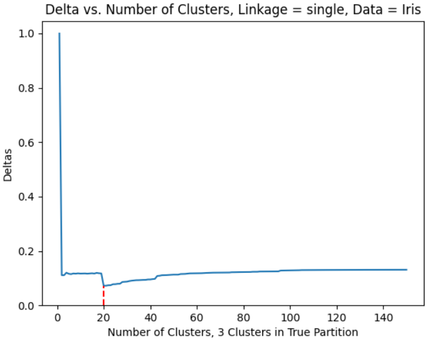


The values of δ plotted against the number of clusters for the *iris* data set using the Ward method. There are 3 clusters in the natural clustering of the set, and the global minimum of δ also occurs at 3 with a value of .049.

o



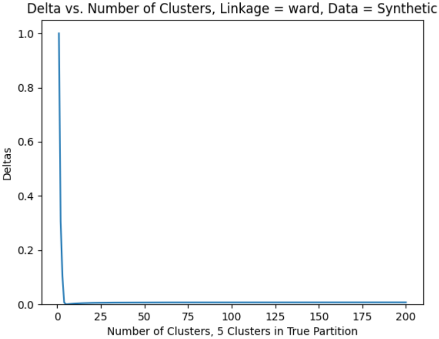
The values of δ plotted against the number of clusters for the *iris* data set using the average link method. There are 3 clusters in the natural clustering of the set, and the global minimum of δ also occurs at 3, with a value of .043.

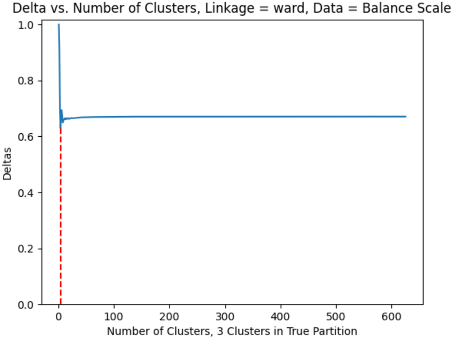


The values of δ plotted against the number of clusters for the *iris* data set using the single link method. The global minimum of δ occurs at 20 with a value of .072. Results are clearly worse than in the previous cases since single-link favors elongated clusters.

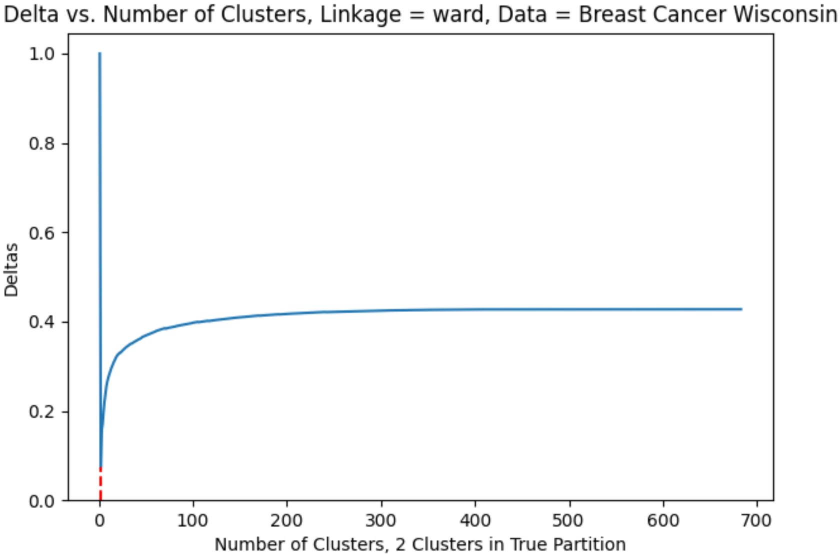
Further tests were performed on a synthetic data set and on several UCI data sets, some of which are notoriously difficult.
Each set was minimally processed by individually normalizing the features to a 0-1 range and removing records with missing values.
After processing, each set was clustered according to the Ward hierarchical method and its δ values computed with its true partition and plotted against number of clusters.

The values of δ plotted against the number of clusters for the synthetic data set using the Ward method. There are 5 clusters in the natural clustering of the set, and the global minimum of δ also occurs at 5, with a value of .0003





Clustering the *balance scale* data set using the Ward method. There are 3 clusters in the natural clustering of the set, but the global minimum of δ occurs at 4 with a value of .630



The values of δ plotted against the number of clusters for the *breast cancer Wisconsin* data set using the Ward method.

There are 2 clusters in the natural clustering of the set, and the global

The results are summarized in the table below.

Data Set	Number of Clusters	minimum δ	Cluster sizes
synthetic	5	.0003	40, 40, 40, 40, 40
iris	3	.049	50, 50, 50
wine	3	.066	59, 71, 48
breast cancer	2	.073	444, 239
zoo	7	.110	41, 20, 5, 13, 4, 8, 10
congressional votes	2	.408	108, 124
glass	6	.624	70, 76, 17, 13, 9, 29
balance scale	3	.630	49, 288, 288

Conclusions

We introduced monotonic entropy and formulated an axiomatization that allows us to extend this type entropy to sets of objects that posses special properties (such as being embedded in a metric space, or being defined by partitions of undirected graphs).

An application of our results is the induction of a metric structure on the set of partitions (clusterings). In turn, this is helpful in the study of stability of clustering algorithms, and for external validation of clusterings, where an apriori data labeling can be compared with the product of a clustering algorithm.

Additionally, by regarding a recommendation system as a bipartite graph between the set of users and set of items, whose edges represent recommendations of items to users, stability of such systems relative to variations in recommendations could be investigated.