ECODAM
Iasi, June 2023

Entropies and Their Applications in Data Mining
Dan A. Simovici
University of Massachusetts Boston
1. The History of the Subject
2. Partitions and Partition Orders
3. Hierarchical Clustering
4. Sum of Square Errors
5. Partition Entropy
6. Axiomatization of Entropy
7. Conditional Entropies and Metrics on Partitions
8. External Validation of Clustering
9. Feature Selection Through Clustering
10. A Greedy Algorithm for Supervised Discretization
11. Further Work and Conclusions
The notion of *entropy* is a probabilistic concept that lies at the foundation of information theory, and originates in thermodynamics and statistical mechanics.

The use of entropy and several related is crucial in many machine learning areas: decision trees, clustering, maximum likelihood algorithms, etc. A powerful generalization of this notion was introduced by Havrda-Charvad and Daroczy, and axiomatized by us.

We define entropies of partitions of sets using an algebraic setting. This approach allows us to take advantage of the partial order that is naturally defined on the set of partitions. Actually, we introduce a generalization of the notion of entropy that has the \textit{Gini index} and \textit{Shannon entropy} as special cases.
Definition

Let $S$ be a nonempty set. A partition of $S$ is a nonempty collection $\pi = \{ B_i \mid i \in I \}$ of nonempty subsets of $S$, such that $\bigcup\{ B_i \mid i \in I \} = S$, and $B_i \cap B_j = \emptyset$ for every $i, j \in I$ such that $i \neq j$.

Each set $B_i$ of $\pi$ is a block of the partition $\pi$.

The set of partitions of a set $S$ is denoted by $\text{PART}(S)$. The partition of $S$ that consists of all singletons of the form \{s\} with $s \in S$ will be denoted by $\alpha_S$; the partition that consists of the set $S$ itself will be denoted by $\omega_S$. 
Example

For the two-element set \( S = \{a, b\} \), there are two partitions: the partition \( \alpha_S = \{\{a\}, \{b\}\} \) and the partition \( \omega_S = \{\{a, b\}\} \).

For the one-element set \( T = \{c\} \), there exists only one partition, \( \alpha_T = \omega_T = \{\{t\}\} \).
Every set of attributes $X$ of a table generates a partition $\pi^X$ on the set of rows, where tuple $t, t'$ belong to the block if $t[X] = t'[X]$.
Example

A complete list of partitions of a set $S = \{a, b, c\}$ consists of

\[
\begin{align*}
\pi_0 &= \{\{a\}, \{b\}, \{c\}\}, \\
\pi_1 &= \{\{a, b\}, \{c\}\}, \\
\pi_2 &= \{\{a\}, \{b, c\}\}, \\
\pi_3 &= \{\{a, c\}, \{b\}\}, \\
\pi_4 &= \{\{a, b, c\}\}.
\end{align*}
\]

Clearly, $\pi_0 = \alpha_S$ and $\pi_4 = \omega_S$. 
Definition

Let $S$ be a set and let $\pi, \sigma \in \text{PART}(S)$. The partition $\pi$ is \textit{finer} than the partition $\sigma$ if every block $C$ of $\sigma$ is a union of blocks of $\pi$. This is denoted by $\pi \leq \sigma$. 
Theorem

Let \( \pi = \{ B_i \mid i \in I \} \) and \( \sigma = \{ C_j \mid j \in J \} \) be two partitions of a set \( S \).

For \( \pi, \sigma \in \text{PART}(S) \), we have \( \pi \preceq \sigma \) if and only if for every block \( B_i \in \pi \) there exists a block \( C_j \in \sigma \) such that \( B_i \subseteq C_j \).
Proof

If $\pi \leq \sigma$, then it is clear for every block $B_i \in \pi$ there exists a block $C_j \in \sigma$ such that $B_i \subseteq C_j$.

Conversely, suppose that for every block $B_i \in \pi$ there exists a block $C_j \in \sigma$ such that $B_i \subseteq C_j$. Since two distinct blocks of $\sigma$ are disjoint, it follows that for any block $B_i$ of $\pi$, the block $C_j$ of $\sigma$ that contains $B_i$ is unique. Therefore, if a block $B$ of $\pi$ intersects a block $C$ of $\sigma$, then $B \subseteq C$.

Let $Q = \bigcup \{B_i \in \pi \mid B_i \subseteq C_j\}$. Clearly, $Q \subseteq C_j$. Suppose that there exists $x \in C_j - Q$. Then, there is a block $B_\ell \in \pi$ such that $x \in B_\ell \cap C_j$, which implies that $B_\ell \subseteq C_j$. This means that $x \in B_\ell \subseteq C$, which contradicts the assumption we made about $x$. Consequently, $C_j = Q$, which concludes the argument.
The order between partitions may express the satisfaction of functional dependencies by tables: $T = (ABCD, R)$ satisfies $AB \rightarrow C$ if and only if $\pi^{AB} \leq \pi^C$ because

$\pi^{AB} = \{\{t_1, t_2\}, \{t_3, t_4, t_5\}, \{t_6, t_7\}\}$, and $\pi^C = \{\{t_1, t_2, t_6, t_7\}, \{t_3, t_4, t_5\}\}$.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>$a_1$</td>
<td>$b_1$</td>
<td>$c_1$</td>
<td>$d_1$</td>
</tr>
<tr>
<td>$t_2$</td>
<td>$a_1$</td>
<td>$b_1$</td>
<td>$c_1$</td>
<td>$d_2$</td>
</tr>
<tr>
<td>$t_3$</td>
<td>$a_1$</td>
<td>$b_2$</td>
<td>$c_2$</td>
<td>$d_1$</td>
</tr>
<tr>
<td>$t_4$</td>
<td>$a_1$</td>
<td>$b_2$</td>
<td>$c_2$</td>
<td>$d_2$</td>
</tr>
<tr>
<td>$t_5$</td>
<td>$a_1$</td>
<td>$b_2$</td>
<td>$c_2$</td>
<td>$d_3$</td>
</tr>
<tr>
<td>$t_6$</td>
<td>$a_2$</td>
<td>$b_2$</td>
<td>$c_1$</td>
<td>$d_1$</td>
</tr>
<tr>
<td>$t_7$</td>
<td>$a_2$</td>
<td>$b_2$</td>
<td>$c_1$</td>
<td>$d_2$</td>
</tr>
</tbody>
</table>
Example

Let $f : S \rightarrow T$ be a function. For $t \in T$ define the set $B_t = \{ x \in S \mid f(x) = t \}$. Then, the collection of sets $\{ B_t \mid t \in T \text{ and } B_t \neq \emptyset \}$ is a partition of $S$ that corresponds to the equivalence $\ker(f)$.

![Diagram showing the relationship between $S$, $T$, $B_t$, and $\pi$](image)
Definition

Let $T$ be a set and let $\pi = \{B_1, \ldots, B_k\} \in \text{PART}(T)$. If $S \subseteq T$, the \textit{trace} of $\pi$ on the set $S$ is the collection of sets:

$$\pi_S = \{B_i \cap S \mid B_i \in \pi \text{ and } B_i \cap S \neq S\}.$$ 

Note that $\pi_S$ is a partition of $S$. 
Example
We have $S = \{5, 6, 7, 12, 13, 14\}$ and $\pi = \{B_1, B_2, B_3, B_4, B_5\}$. The trace of $\pi$ on $S$ denoted by $\pi_S$ consists of

\[
\begin{align*}
B_2 \cap S &= \{12\}, \\
B_3 \cap S &= \{13, 14\}, \\
B_5 \cap S &= \{5, 6, 7\}.
\end{align*}
\]
On slide 9 we introduced the partial order $\leq$ on the set $\text{PART}(S)$.

**Definition**

Let $\pi, \sigma \in \text{PART}(S)$ be two partitions of the finite set $S$. We say that $\sigma$ covers $\pi$ (and write $\pi \triangleleft \sigma$) if $\pi = \{B_1, \ldots, B_m\}$ and the blocks of $\sigma$ are the same as the blocks of $\pi$, except for a block $C$ of $\sigma$ that is the union of two blocks $B', B''$ of $\pi$.

Note that if $\pi \triangleleft \sigma$ and $\pi$ has $m$ blocks, then $\sigma$ has $m - 1$ blocks.
Example

Let $S = \{1, 2, 3, 4, 5, 6\}$ and let $\pi = \{\{1\}, \{2, 4\}, \{3, 5, 6\}\}$. Then, the partitions

$$\begin{align*}
\sigma_1 &= \{\{1, 2, 4\}, \{3, 5, 6\}\}, \\
\sigma_2 &= \{\{1, 3, 5, 6\}, \{2, 4\}\}, \\
\sigma_3 &= \{\{1\}, \{2, 3, 3, 5, 6\}\}
\end{align*}$$

all cover $\pi$. 
If \( \pi \leq \sigma \) for \( \pi, \sigma \in \text{PART}(S) \) there exists a sequence of partitions \( \tau_1, \ldots, \tau_k \) such that

\[
\pi = \tau_1 \triangleleft \tau_2 \triangleleft \cdots \triangleleft \tau_k = \sigma.
\]

In other words, if \( \pi \leq \sigma \) it is possible to “interpolate” the partitions \( \tau_1, \ldots, \tau_k \) such that each partition \( \tau_\ell \) is covered by \( \tau_{\ell+1} \) for \( 1 \leq \ell \leq k - 1 \).
Example

Let $\pi = \{\{1, 2\}, \{3\}, \{4, 5, 6\}, \{7, 8\}, \{9\}\}$ and $\sigma = \{\{1, 2, 3, 7, 8\}, \{4, 5, 6, 9\}\}$. Clearly, $\pi \leq \sigma$.

The chain of partitions $\tau_1 \triangleleft \tau_2 \triangleleft \tau_3 \triangleleft \tau_4$ can be interpolated between $\pi$ and $\sigma$, where

$$\tau_1 = \{\{1, 2\}, \{3\}, \{4, 5, 6\}, \{7, 8\}, \{9\}\} = \pi,$$
$$\tau_2 = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8\}, \{9\}\},$$
$$\tau_3 = \{\{1, 2, 3\}, \{4, 5, 6, 9\}, \{7, 8\}\}$$
$$\tau_4 = \{\{1, 2, 3, 7, 8\}, \{4, 5, 6, 9\}\} = \sigma.$$
Definition

Let $S$, $T$ be two disjoint sets and let $\sigma = \{B_1, \ldots, B_m\} \in \text{PART}(S)$ and $\tau = \{C_1, \ldots, C_n\} \in \text{PART}(T)$. The sum of the partitions $\sigma$ and $\tau$ is the partition $\sigma + \tau$ of the set $S \cup T$ given by:

$$\sigma + \tau = \{B_1, \ldots, B_m, C_1, \ldots, C_n\}.$$
\[ \pi = \{ B_1, B_2, B_3 \} \in \text{PART}(S), \sigma = \{ C_1, C_2, C_3, C_4 \} \in \text{PART}(T) \]
\[ \pi + \sigma = \{ B_1, B_2, B_3, C_1, C_2, C_3, C_4 \} \in \text{PART}(S \cup T). \]
Intersection of Two Partitions

Definition

Let now $\pi, \tau$ be two partitions in $\text{PART}(S)$, where

$$\pi = \{B_1, \ldots, B_m\},$$
$$\tau = \{D_1, \ldots, D_p\},$$

The partition $\pi \land \tau \in \text{PART}(S)$ is

$$\pi \land \tau = \{B_i \cap D_j \mid B_i \cap D_j \neq \emptyset\}.$$
Example

Let \( S = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\} \) and let

\[
\pi = \{\{x_1\}, \{x_2, x_3, x_4, x_5, x_6\}, \{x_7\}\} \\
\tau = \{\{x_1, x_2, x_3\}, \{x_4, x_5, x_6, x_7\}\}.
\]

We have

\[
\pi \land \tau = \{\{x_1\}, \{x_2, x_3\}, \{x_4, x_5, x_6\}, \{x_7\}\}.
\]
What is Hierarchical Clustering?

- A **hierarchical clustering algorithm** consists in constructing a chain of partitions of the set of objects referred usually as a *dendrogram*.

- The blocks of the initial partition of the set of objects \( S \) are singletons, that is, the initial partition of \( S \) is

\[
\alpha_S = \{ \{x\} \mid x \in S \}.
\]

- Starting from a given partition of \( S \), a new partition is constructed by fusing a pair of blocks of the previous partition in the chain. The blocks to be fused are selected according to a criterion specific to the clustering method.
A variety of methods have been developed for constructing such clusterings (single-link, complete-link, Ward clusterings, group-average, centroid clusterings).

Each of these methods facilitates finding certain types of clusters (elongated clusters in the case of single-link, globular clusters in the case of complete-link, etc.)
Clustering sets of vectors in $\mathbb{R}^P$ amounts to partitioning these sets such that the resulting partitions (referred to as *clusterings*) satisfy certain desirable properties. Depending on specific circumstances we may insist that clusters are well-separated, or have approximate equal sizes, or have other properties.
A Simple Example

Let \( S = \{ x_1, x_2, x_3, x_4 \} \) consist of

\[
x_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad x_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad x_4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},
\]

\[
x_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad x_4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}
\]
Hierarchical Clustering
Any path in this diagram from $\{\{1\}, \{2\}, \{3\}, \{4\}\}$ to $\{1, 2, 3, 4\}$ is a clustering.
An equivalent representation of this path is the dendrogram.
There are several methods to fuse clusterings that yield very distinct results.

The next table contains the cluster dissimilarity that is minimized when two clusters $U$ an $V$ are fused in some of the most popular hierarchical clustering methods.

<table>
<thead>
<tr>
<th>Clustering method</th>
<th>Cluster Dissimilarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>single link</td>
<td>$sl(U, V) = \min{d(u, v) \mid u \in U, v \in V}$</td>
</tr>
<tr>
<td>complete link</td>
<td>$cl(U, V) = \max{d(u, v) \mid u \in U, v \in V}$</td>
</tr>
<tr>
<td>group average</td>
<td>$gav(U, V) = \frac{\sum{d(u, v) \mid u \in U, v \in V}}{</td>
</tr>
<tr>
<td>centroid method</td>
<td>$cen(U, V) = |\mathbf{c}_U - \mathbf{c}_V|^2$</td>
</tr>
<tr>
<td>Ward method</td>
<td>$ward(U, V) = \frac{</td>
</tr>
</tbody>
</table>
The thick lines represent clusterings achieved with the Ward method.
The Sum-of-Squares

The notion of *inertia* of a finite subset $S$ of $\mathbb{R}^p$ relative to a vector $z$ originates in mechanics of solids. The *inertia of $S$ relative to a vector $z \in \mathbb{R}^p$* is the number

$$I_z(S) = \sum_{x \in S} \| x_j - z \|_2^2.$$  

The *centroid* of a finite set of vectors $\{x_1, \ldots, x_m\}$ is defined as:

$$c_S = \frac{1}{|S|} \sum_{i=1}^{m} x_i.$$
Huygens’ Inertia Theorem stipulates that

$$l_z(S) - l_{c_S}(S) = m \| c_S - z \|^2,$$

for every $z \in \mathbb{R}^p$. Thus, the minimal value of the inertia $l_z(S)$ is achieved for $z = c_S$.

The special case of the inertia of $S$ relative to the centroid $c_S$ is referred to as the \textit{sum of square errors}. We denote $l_{c_X}(S)$ by $\text{sse}(S)$.

The sum of square errors is a measure of sets of objects that can be defined when the objects belong to a metric space (e.g. $\mathbb{R}^n$).
Properties of \( \text{sse} \)

- If \( W \subseteq S \) and \( \sigma = \{ U, V \} \) be a bipartition of \( W \). We have:

\[
\text{sse}(W) = \text{sse}(U) + \text{sse}(V) + \frac{|U||V|}{|W|} \| c_U - c_V \|_2^2 .
\]

- If \( U \subseteq W \), then \( \text{sse}(U) \leq \text{sse}(W) \).
The notion of entropy is a probabilistic concept that lies at the foundation of information theory.

Our goal is to define entropy in an algebraic setting by introducing the notion of entropy of a partition of a finite set.

This approach allows us to take advantage of the partial order that is naturally defined on the set of partitions.
Actually we introduce two distinct type of entropies:

- **cardinality entropies** that are defined using the cardinalities of partition blocks, and
- **inertial entropies** that use the metric (distance) between objects.

The second type of entropies involves new, recent results. Cardinality entropies will be denote by $\mathcal{H}_\beta$, where $\beta > 1$; inertial entropy is denoted by $\mathcal{H}_s$.

When statements are valid for both types of entropies we’ll use $\mathcal{H}$. 
In classical information theory the *Shannon entropy* of a probability distribution \( p = (p_1, \ldots, p_m) \), where \( p_i > 0 \) for \( 1 \leq i \leq m \) and \( p_1 + \cdots + p_m = 1 \) is defined as

\[
\mathcal{H}(p_1, \ldots, p_m) = - \sum_{i=1}^{m} p_i \log_2 p_i. = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i}.
\]

If \( \pi = \{B_1, \ldots, B_m\} \) is a partition of a set \( S \), then a probability distribution \( p_\pi \) can be defined as

\[
p_\pi = \left( \frac{|B_1|}{|S|}, \ldots, \frac{|B_m|}{|S|} \right).
\]

Accordingly, we can define the *Shannon entropy of a partition* \( \pi \) as:

\[
\mathcal{H}(\pi) = - \sum_{i=1}^{m} \frac{|B_i|}{|S|} \log_2 \frac{|B_i|}{|S|}.
\]
Example

Let $S$ be a set containing ten elements and let $\pi_1, \pi_2, \pi_3, \pi_4$ be the four partitions shown below.

- $\mathcal{H}(\pi_1) = 2.32$
- $\mathcal{H}(\pi_2) = 2.17$
- $\mathcal{H}(\pi_3) = 2.04$
- $\mathcal{H}(\pi_4) = 1.96$
The partition $\pi_1$, which is the most uniform (each block containing two elements), has the largest entropy. At the other end of the range, partition $\pi_4$ has a strong concentration of elements in its fourth block and the lowest entropy.

The entropy can be viewed as a measure of impurity of a partition.
Definition

The *Gini index* of $\pi$ is the number

$$gini(\pi) = 1 - \sum_{i=1}^{m} \left( \frac{|B_i|}{|S|} \right)^2.$$

Like Shannon entropy, the Gini index can be used to evaluate the uniformity of the distribution of the elements of $S$ in the blocks of $\pi$ because both $H(\pi)$ and $gini(\pi)$ increase with the uniformity of the distribution of the elements of $S$. 
Example

Results concerning the Gini index are shown next:

\[
gini(\pi_1) = 0.80
\]

\[
gini(\pi_2) = 0.79
\]

\[
gini(\pi_3) = 0.72
\]

\[
gini(\pi_4) = 0.68
\]
Generalized Entropy

**Definition**

Let \( \pi = \{B_1, \ldots, B_m\} \) be a partition of a set \( S \) and let \( \beta > 1 \). The \( \beta \)-entropy of a partition \( \pi \) is the number

\[
\mathcal{H}_\beta(\pi) = \frac{1}{1 - 2^{1-\beta}} \cdot \left( 1 - \sum_{i=1}^{m} \left( \frac{|B_i|}{|S|} \right)^\beta \right).
\]

If \( \beta = 2 \), we obtain \( \mathcal{H}_2(\pi) \), which is twice the Gini index,

\[
\mathcal{H}_\beta(S, \pi) = 2 \cdot \left( 1 - \sum_{i=1}^{m} \left( \frac{|B_i|}{|S|} \right)^2 \right).
\]

The *Gini index*, \( \text{gini}(\pi) = 1 - \sum_{i=1}^{m} \left( \frac{|B_i|}{|S|} \right)^2 \), is widely used in machine learning and data mining.
When we take \( \lim_{\beta \to 1} H_{\beta}(\pi) \) we obtain the Shannon entropy!
Indeed, we can write:

\[
\lim_{\beta \to 1} H_{\beta}(\pi) = \lim_{\beta \to 1} \frac{1}{1 - 2^{1-\beta}} \cdot \left(1 - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta\right)
\]

\[
= \lim_{\beta \to 1} - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta \ln \frac{|B_i|}{|S|}
\]

(by l’Hôpital Rule)

\[
= - \sum_{i=1}^{m} \frac{|B_i|}{|S|} \log_2 \frac{|B_i|}{|S|}.
\]
Lemma

If $\pi, \sigma \in \text{PART}(S)$ and $\pi \prec \sigma$, then $\mathcal{H}_\beta(\pi) \geq \mathcal{H}_\beta(\sigma)$.

Proof: Since $\pi \prec \sigma$ we can write

$$\pi = \{B_1, \ldots, B_m\} \text{ and } \sigma = \{B_1, \ldots, B_{m-2}, B_{m-1} \cup B_m\},$$

and the inequality $\mathcal{H}(\pi) \geq \mathcal{H}(\sigma)$ follows from the fact that

$$|B_{m-1}|^\beta + |B_m|^\beta \leq |B_{m-1} \cup B_m|^\beta \text{ for } \beta \geq 1.$$
We proposed the following set of axioms for the $\mathcal{H}_\beta$ entropy:

(P1) If $\pi, \pi' \in \text{PART}(A)$ are such that $\pi \leq \pi'$, then $\mathcal{H}_\beta(\pi') \leq \mathcal{H}_\beta(\pi)$.

(P2) If $A$ and $B$ are two finite sets such that $|A| \leq |B|$, then $\mathcal{H}_\beta(\iota_A) \leq \mathcal{H}_\beta(\iota_B)$.

(P3) For every disjoint sets $A, B$ and partitions $\pi \in \text{PART}(A)$, and $\sigma \in \text{PART}(B)$ we have:

$$\mathcal{H}_\beta(\pi + \sigma) = \left(\frac{|A|}{|A| + |B|}\right)^\beta \mathcal{H}_\beta(\pi) + \left(\frac{|B|}{|A| + |B|}\right)^\beta \mathcal{H}_\beta(\sigma) + \mathcal{H}_\beta(\{A, B\}).$$

(P4) If $\pi \in \text{PART}(A)$ and $\sigma \in \text{PART}(B)$, the

$$\mathcal{H}_\beta(\pi \times \sigma) = \Phi(\mathcal{H}_\beta(\pi), \mathcal{H}_\beta(\sigma)),$$

where $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a continuous function such that $\Phi(x, y) = \Phi(y, x)$, and $\Phi(x, 0) = x$ for $x, y \in \mathbb{R}$. 
It was shown that a function $\mathcal{H}_\beta$ that satisfies the axioms (P1)–(P4) and for any partition $\pi = \{B_1, \ldots, B_m\} \in \text{PART}(S)$ has necessarily the form:

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

When $\beta = 2$, $\mathcal{H}_2(\pi) = 2\left(\sum_{i=1}^{m} \left( \frac{|B_i|}{|S|} \right)^2 - 1\right)$, which is twice the value of the Gini index. On other hand, we have

$$
\lim_{\beta \to 1, \beta > 1} \mathcal{H}_\beta(\pi) = -\sum_{i=1}^{m} \frac{|B_i|}{|S|} \log_2 \frac{|B_i|}{|S|},
$$

which corresponds to Shannon entropy.

The $\beta$-entropy of a partition depends exclusively on the cardinalities of the blocks of the partition.
Another type of entropy: inertial entropy

When the set $S$ is a subset of a metric space, it becomes possible to define an entropy-like measure that is defined by the scattering of the elements of blocks around the centroids of these blocks, which supplements the description of partitions.
Axiomatization of inertial entropy

Let $\mathcal{P}_{\text{fin}}(\mathbb{R}^p)$ be the collection of finite subsets of $\mathbb{R}^p$. We introduce the inertial entropy of a partition of a finite subset $S$ of $\mathbb{R}^p$ as a function $\mathcal{H}_s : \mathcal{P}_{\text{fin}}(\mathbb{R}^p) \rightarrow \mathbb{R}_{\geq 0}$ that satisfies the following axioms:

- **Initial Values Axiom:** $\mathcal{H}_s(\alpha S) = 1$ and $\mathcal{H}_s(\omega S) = 0$.

- **Addition Axiom:** If $S$ and $T$ are finite disjoint sets, $\pi \in \text{PART}(S)$, and $\sigma \in \text{PART}(T)$ then

  \[
  \mathcal{H}_s(\pi + \sigma) = \frac{\text{sse}(S)}{\text{sse}(S \cup T)} \mathcal{H}_s(\pi) + \frac{\text{sse}(T)}{\text{sse}(S \cup T)} \mathcal{H}_s(\sigma) + \mathcal{H}_s(\{S, T\}).
  \]

Note the similarity between the Addition Axiom for inertial entropy and the axiom P3 for partition entropy.
The axioms of inertial entropy suffice for defining it:

Theorem

If $\pi = \{B_1, \ldots, B_m\}$ be a partition of a finite subset $S$ of $\mathbb{R}^p$, then

$$\mathcal{H}_s(\pi) = 1 - \frac{1}{\text{sse}(S)} \sum_{i=1}^{m} \text{sse}(B_i) = 1 - \frac{\text{sse}(\pi)}{\text{sse}(\omega_S)}.$$
Many properties are shared between $\beta$-entropy and inertial entropy.

- Recall that $\beta$-entropy is denoted by $\mathcal{H}_\beta$; the inertial entropy is denoted by $\mathcal{H}_s$.
- When properties are shared by $\beta$-entropy and by inertial entropy we denote either one just by $\mathcal{H}$. 
Definition

Let \( \pi, \sigma \in \text{PART}(S) \) and let \( \sigma = \{C_1, \ldots, C_n\} \).

The \( \beta \)-conditional entropy is the function \( \mathcal{H}_\beta : \text{PART}(S)^2 \rightarrow \mathbb{R}_{\geq 0} \) defined by

\[
\mathcal{H}_\beta(\pi | \sigma) = \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^\beta \mathcal{H}_\beta(\pi_{C_j}).
\]

The inertial conditional entropy is \( \mathcal{H}_s : \text{PART}(S)^2 \rightarrow \mathbb{R}_{\geq 0} \) defined as

\[
\mathcal{H}_s(\pi | \sigma) = \sum_{j=1}^{n} \frac{sse(C_j)}{sse(S)} \mathcal{H}_s(\pi_{C_j}).
\]
Note that $\mathcal{H}_s(\pi|\omega_S) = \mathcal{H}_s(\pi)$,

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

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

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

$$\mathcal{H}_s(\pi | \sigma) = \sum_{j=1}^{n} \frac{sse(C_j)}{sse(S)} \mathcal{H}(\pi C_j)$$

$$= \sum_{j=1}^{n} \frac{sse(C_j)}{sse(S)} \left( 1 - \frac{1}{sse(C_j)} \sum_{i=1}^{m} sse(B_i \cap C_j) \right)$$

$$= \frac{1}{sse(S)} \sum_{j=1}^{n} \left( sse(C_j) - \sum_{i=1}^{n} sse(B_i \cap C_j) \right)$$

$$= \mathcal{H}_s(\pi \land \sigma) - \mathcal{H}_s(\sigma).$$
Note that for $\pi \in \text{PART}(S)$ we have:

$$\mathcal{H}_\beta(\pi | \omega S) = \mathcal{H}_\beta(\pi)$$

and that

$$\mathcal{H}_\beta(\omega S | \sigma) = \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^\beta \mathcal{H}_\beta(\omega_{C_j}) = 0,$$

$$\mathcal{H}_\beta(\pi | \alpha S) = \sum_{j=1}^{n} \frac{1}{|S|} \mathcal{H}(\pi_{\{x_j\}}) = 0$$

for every partition $\pi \in \text{PART}(S)$. 
For $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$, the conditional entropy can be written explicitly as

$$
\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^\beta \sum_{i=1}^{m} \frac{1}{1 - 2^{1-\beta}} \left[ 1 - \left( \frac{|B_i \cap C_j|}{|C_j|} \right)^\beta \right]
$$

$$
= \frac{1}{1 - 2^{1-\beta}} \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^\beta - \sum_{i=1}^{m} \left( \frac{|B_i \cap C_j|}{|S|} \right)^\beta \tag{1}
$$

For the special case when $\pi = \alpha_S$, we can write

$$
\mathcal{H}_\beta(\alpha_S|\sigma) = \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^\beta \mathcal{H}_\beta(\alpha_{C_j}) = \frac{1}{1 - 2^{1-\beta}} \left( \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^\beta \right) - \frac{1}{|S|^{\beta-1}} \tag{2}
$$
Theorem

(Anti-monotonicity of Entropy) If $\pi, \sigma \in \text{PART}(S)$ and $\pi \leq \sigma$, then $\mathcal{H}(\pi) \geq \mathcal{H}(\sigma)$.

Proof: Since $\pi \leq \sigma$ for $\pi, \sigma \in \text{PART}(S)$ there exists an interpolating sequence of partitions $\tau_1, \ldots, \tau_k$ such that

$$\pi = \tau_1 \triangleleft \tau_2 \triangleleft \cdots \triangleleft \tau_k = \sigma.$$

By the previous lemma we have $\mathcal{H}(\tau_1) \leq \mathcal{H}(\tau_2) \leq \cdots \leq \mathcal{H}(\tau_k)$, hence $\mathcal{H}(\pi) \geq \mathcal{H}(\sigma)$.
This is the anti-monotonicity property of entropy.
The next statement is a generalization of a well-known property of Shannon’s entropy.

**Theorem**

Let $\pi$ and $\sigma$ be two partitions of a finite set $S$. We have

$$H(\pi \wedge \sigma) = H(\pi|\sigma) + H(\sigma) = H(\sigma|\pi) + H(\pi),$$
Theorem

The mapping \( d : \text{PART}(S)^2 \rightarrow \mathbb{R}_{\geq 0} \) defined as

\[
d(\pi, \sigma) = \mathcal{H}(\pi | \sigma) + \mathcal{H}(\sigma | \pi)
\]

is a metric on \( \text{PART}(S) \).

This generates a family of metrics \( \{d_\beta \mid \beta \geq 1\} \); in addition, we obtain the inertial metric \( d_s \) derived from the inertial entropy.
Proof

Let \( \pi = \{B_1, \ldots, B_m\} \) and \( \sigma = \{C_1, \ldots, C_n\} \) be two partitions of \( S \).

We have

\[
\mathcal{H}_\beta(\pi \land \sigma) - \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^\beta \mathcal{H}(\pi_{C_j})
\]

\[
= \frac{1}{1 - 2^{1-\beta}} \left( 1 - \sum_{i} \sum_{j} \left( \frac{|B_i \cap C_j|}{|S|} \right)^\beta \right)
\]

\[
- \frac{1}{1 - 2^{1-\beta}} \sum_{j} \left( \frac{|C_j|}{|S|} \right)^\beta \left( 1 - \sum_{i} \left( \frac{|B_i \cap C_j|}{|C_j|} \right)^\beta \right)
\]

\[
= \mathcal{H}_\beta(\sigma).
\]
From the result established above

\[ H_\beta(\pi \land \sigma) = \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^\beta H_\beta(\pi_{C_j}) + H_\beta(\sigma), \]

we obtain

\[ H_\beta(\pi \land \sigma) = H_\beta(\pi \mid \sigma) + H_\beta(\sigma). \]

The second equality has a similar proof.
### Definition

Let $\pi, \sigma$ be two partitions of a set $S$, where $\sigma = \{C_1, \ldots, C_n\}$. The $\beta$-information gain of $\sigma$ is the number

$$\text{gain}(\pi, \sigma) = H_\beta(\pi) - H_\beta(\pi | \sigma).$$
Let $\pi \in \text{PART}(S)$, and let $\sigma = \{C_1, \ldots, C_n\} \in \text{PART}(S)$. If $\beta \to 1$ the information gain of the Shannon entropy is

$$\text{gain}(\pi, \sigma) = H_\beta(\pi) - H_\beta(\pi|\sigma)$$

$$= H_\beta(\pi) - \sum_{i=1}^{n} \frac{|C_j|}{|S|} H_\beta(\pi_{C_j}).$$
Note that $\pi \succeq \sigma$, where $\pi, \sigma \in \text{PART}(S)$, we have

$$\mathcal{H}_\beta(\pi|\sigma) = \mathcal{H}_\beta(\pi)$$

if and only if

$$\text{gain}(\pi, \sigma) = \mathcal{H}_\beta(\pi) - \mathcal{H}_\beta(\pi|\sigma) = 0,$$

by the Theorem on slide 59.
Theorem

Let $\pi, \sigma \in \text{PART}(S)$, where $S$ is a finite set. We have $\mathcal{H}(\pi|\sigma) = 0$ if and only if $\sigma \leq \pi$. 
Corollary

If \( \pi = \{B_1, \ldots, B_m\} \) and \( \sigma = \{C_1, \ldots, C_n\} \) are two partitions of the set \( S \), then

\[
d_{\beta}(\pi, \sigma) = \frac{1}{1 - 2^{1-\beta}} \left( \sum_{i=1}^{m} \left( \frac{|B_i|}{|S|} \right)^{\beta} + \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^{\beta} \right)
- 2 \sum_{i=1}^{m} \sum_{j=1}^{n} \left( \frac{|B_i \cap C_j|}{|S|} \right)^{\beta} \right)
\]
Taking $\beta = 2$ we obtain the distance

$$d_2(\pi, \sigma) = 2 \left( \sum_{i=1}^{m} \left( \frac{|B_i|}{|S|} \right)^2 + \sum_{j=1}^{n} \left( \frac{|C_j|}{|S|} \right)^2 - 2 \sum_{i=1}^{m} \sum_{j=1}^{n} \left( \frac{|B_i \cap C_j|}{|S|} \right)^2 \right)$$

$$= \frac{2}{|S|^2} \left( \sum_{i=1}^{m} |B_i|^2 + \sum_{j=1}^{n} |C_j|^2 - 2 \sum_{i=1}^{m} \sum_{j=1}^{n} |B_i \cap C_j|^2 \right). \quad (3)$$

Note that $d_\beta(\alpha_S, \pi) = H_\beta(\alpha_S) - H_\beta(\pi)$ and $d_\beta(\omega_S, \pi) = H_\beta(\pi)$. 
The metric $d_\beta$ is normalized, by defining the mapping 
$e_\beta : \text{PART}(S)^2 \longrightarrow \mathbb{R}^{\geq 0}$ as

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

and $0 \leq e_\beta(\pi, \sigma) \leq 1$ for every $\beta \geq 1$.

Similarly, the mapping $\delta : \text{PART}(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ derived from the inertial entropy and defined as

$$\delta(\pi, \sigma) = \frac{d_s(\pi, \sigma)}{\mathcal{H}(\pi \wedge \sigma)}$$

is a metric on $\text{PART}(S)$ such that $0 \leq \delta(\pi, \sigma) \leq 1$ for $\pi, \sigma \in \text{PART}(S)$. 
It is interesting that a special case, $d_2(\pi, \beta)$, that is, the metric that corresponds to the Gini entropy has been introduced previously starting from combinatorial considerations and is also known as the Barthelemy-Montjardet metric.

Namely, for a set $S$ and two partitions $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$, the number of unordered pairs $\{x, y\}$ of elements of $S$ that belong to the same block of $\pi$ is

$$\sum_{i=1}^{m} \binom{|B_i|}{2} = \sum_{i=1}^{m} \frac{1}{2}(|B_i|^2 - |B_i|) = \frac{1}{2} \sum_{i=1}^{m} |B_i|^2 - \frac{1}{2} |S|.$$ 

The similar number for $\sigma$ is $\frac{1}{2} \sum_{j=1}^{n} |C_j|^2 - \frac{1}{2} |S|$. For the partition $\pi \land \sigma$, the number of pairs that belong to the same block of both $\pi$ and $\sigma$ is $\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} |B_i \cap C_j|^2 - \frac{1}{2} |S|$.
Thus, the number of pairs that belong to exactly one block of either partition is

\[
\frac{1}{2} \sum_{i=1}^{m} |B_i|^2 + \frac{1}{2} \sum_{j=1}^{n} |C_j|^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} |B_i \cap C_j|^2
\]

\[
= \frac{1}{2} \left( \sum_{i=1}^{m} |B_i|^2 + \sum_{j=1}^{n} |C_j|^2 - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} |B_i \cap C_j|^2 \right) = \frac{1}{4} d_2(\pi, \sigma),
\]

which shows that the Barthélemy-Montjardet metric is proportional with \(d_2(\pi, \sigma)\).
Example

Let \( S = \{x_1, \ldots, x_{12}\} \) be a set and let \( \pi, \sigma \) be the partitions defined as

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

and

\[ \sigma = \{\{x_1, x_7, x_9\}, \{x_2, x_3, x_4\}, \{x_5, x_6, x_8, x_{10}\}\}. \]

These partition are represented by vectors of dimension \(|S|\) that indicate the number of the block to which successive elements belong:

\[ \mathbf{v}_\pi = (1, 2, 3, 1, 2, 4, 1, 3, 4, 4), \]
\[ \mathbf{v}_\sigma = (1, 2, 2, 2, 3, 3, 1, 3, 1, 3). \]
If $m_{ij}$ is the $(i,j)$-entry of the contingency matrix $M_{\pi,\sigma}$, the Pearson $\chi^2_{\pi,\sigma}$ association index of $\pi$ and $\sigma$ can be written in our framework as

$$\chi^2_{\pi,\sigma} = \sum_i \sum_j \frac{(p_{ij} - |B_i||C_j|)^2}{|B_i||C_j|}.$$ (4)

It is well-known that the asymptotic distribution of $\chi^2_{\pi,\sigma}$ is a $\chi^2$-distribution with $|\pi||\sigma|$ degrees of freedom.
External clustering validation involves clustering data that is labeled.

The assumption is that the labels denote the correct cluster where a data item belongs and this define a *ground truth partition* $\sigma$.

A clustering algorithm produces a partition $\pi$, and the goal of this type of validation is to determine to what extent the partition $\pi$ is consistent with the ground truth partition $\sigma$. 
The consistency is evaluated by using the normalized distance $\delta(\pi, \sigma) \in [0, 1]$. The smaller values of $\delta$ (close to 0) mean that the clustering algorithm produces a result consistent with the ground truth partition.

As a secondary effect, 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 $\pi$ are variable.
We use two data sets with known clusterings:

- the first is a synthetic data set consisting of five randomly generated Gaussian clusters in $\mathbb{R}^2$,
- the second is the *iris* data set which is known to have three true classes. In both cases, the data sets are clustered according to the Ward hierarchical method.
Once the clustering is completed, for each partition $\pi$ in the resulting dendrogram, we compute the value of $\delta(\pi, \sigma)$ where $\sigma$ is the ground truth partition. The values of $\delta(\pi, \sigma)$ are then plotted against the number of clusters to visualize the behavior of the $\delta$ function.
Variation of $\delta(\pi, \sigma)$ plotted against the number of clusters for a synthetic data set that consists of five clusters.

Minimum of $\delta(\rho, \sigma)$ at $n = 5$ for the synthetic data set.
Variation of $\delta(\pi, \sigma)$ plotted against the number of clusters for the iris data set that consists of three clusters using the Ward method.

A zoomed view of the local minimum of the left figure.
We also investigated the effect that choice of hierarchical clustering method has on the effectiveness of the $\delta$ function at partition validation. To do so, we repeat the same experiment as above but change the hierarchical method in each new plot set.
Variation of $\delta(\pi, \sigma)$ plotted against the number of clusters for the iris data set that consists of three clusters using the average link method.

A zoomed view of the local minimum of the left figure.
External Validation of Clustering

Variation of $\delta(\pi, \sigma)$ plotted against the number of clusters for the iris data set that consists of three clusters using the complete link method.

A zoomed view of the local minimum of the left figure.
Variation of $\delta(\pi, \sigma)$ plotted against the number of clusters for the iris data set that consists of three clusters using the single link method.

A zoomed view of the local minimum of the left figure.
The average and complete link methods are comparable to the Ward method’s results. The three methods all have a global minimum at the correct number of clusters.

The single link method is the worst choice. It has a minimum distance to the real partition at around 20 clusters, which is very different from the label set’s 3. This can be partially explained by the single link method propensity to prefer clusters that have elongated shapes.
The performance, robustness, and usefulness of classification algorithms are improved when relative few features are involved in the classification.

Several approaches to feature selection have been explored including wrapper techniques, support vector machines, neural networks, and prototype-based feature selection that is closed to our approach.
In previous work we presented an algorithm for feature selection that clusters attributes using an entropic metric and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. The main interest of this approach resides in the improved understanding of the structure of the analyzed data and of the relative importance of the attributes for the selection process.
A metric $\delta(A, B)$ on the set of attributes of a data set was defined starting with the metric between the corresponding partitions $\pi^A, \pi^B$. If we use the Gini metric, $\delta$ is defined as

$$
\delta(\pi^A, \pi^B) = \sum_i |U_i|^2 + \sum_j |V_j|^2 - 2 \sum_i \sum_j |U_i \cap V_j|^2,
$$

where $\pi^A = \{U_1, \ldots, U_m\}$, and $\pi^B = \{V_1, \ldots, V_n\}$. The contingency matrix of these partitions is the matrix $M_{\pi^A, \pi^B}$. 
The Pearson coefficient $\chi^2_{\pi,\sigma}$ decreases with the distance between partitions and, thus, the probability that $\pi$ and $\sigma$ are independent increases with this distance.

This suggests that partitions that are correlated are close in the sense of the Barthélemy-Montjardet metric. Therefore, if the attributes of the data set are clustered using the $\delta$-distance between partitions we could replace clusters with their medoids and, thereby, drastically reduce the number of attributes involved in a classification without significant decreases in accuracy of the resulting classifiers.
Experimental results justify this claim.

- An experiment conducted on the data set votes which records the votes of 435 US congressman on 15 key questions (and each congressman is classified as a democrat or republican) shows that retaining only 7 key attributes as representatives of clusters obtained using the Ward clustering method does not result in an appreciable loss of accuracy.

- A classification produced with a wrapper with the J48 algorithm results in an accuracy of 96.03%.
Frequently data sets have attributes with numerical domains which makes them unsuitable for certain data mining algorithms that deal mainly with nominal attributes, such as decision trees and naive Bayes classifiers. To use such algorithms we need to replace numerical attributes with nominal attributes that represent intervals of numerical domains with discrete values. This process, known to as *discretization*, has received a great deal of attention in the data mining literature and includes a variety of ideas:

- fixed $k$-interval discretization,
- fuzzy discretization,
- Shannon-entropy discretization due to Fayyad and Irani,
- proportional $k$-interval discretization,
- techniques that are capable of dealing with highly dependent attributes.
To discretize a numerical attribute $B$ we select a sequence of numbers $t_1 < t_2 < \cdots < t_\ell$ in $\text{adom}(B)$. Next, the attribute $B$ is replaced by the nominal attribute $\hat{B}$ that has $\ell + 1$ distinct values in its active domain $\{k_0, k_1, \ldots, k_\ell\}$. Each $B$-component $b$ of an object $o$ is replaced by the discretized $B$-component $k$ defined by

$$
k = \begin{cases} 
k_0 & \text{if } b \leq t_1, \\
k_i & \text{if } t_i < b \leq t_{i+1} \text{ for } 1 \leq i \leq \ell - 1, \\
k_\ell & \text{if } t_\ell < b. \end{cases}$$
The starting point of our result is the observation that every nominal attribute $A$ of a set of objects $S$ induces a partition $\kappa_A$ of the set $S$ such that the objects $t, s$ belong to the same block of the partition $\kappa_A$ if their $A$-components are equal. SQL computes the partition $\kappa_A$ using the group by option of a select phrase.
There are two types of discretization:

- **unsupervised discretization**, where the discretization takes place without any knowledge of the classes to which objects belong, and

- **supervised discretization** which takes into account the classes of the objects.

Our approach involves **supervised discretization**. Within our framework, to discretize a numerical attribute $B$ amounts to constructing a partition of the active domain $\text{adom}(B)$ taking into account the partition $\kappa_A$ determined by the nominal class attribute $A$. 
The set of numbers
\[ T = (t_1, t_2, \ldots, t_\ell) \]
defines the discretization process and its members are referred to as **class separators**.

The corresponding partition of \( \text{adom}(B) \) is \( \pi_B^T = \{ Q_0, \ldots, Q_\ell \} \), where \( Q_i = \{ b \in \text{Dom}(B) \mid t_i \leq b \leq t_{i+1} \} \) for \( 0 \leq i \leq \ell \), where \( t_0 = -\infty \) and \( t_{\ell+1} = \infty \).

It is immediate that \( \pi_B^{T \cup T'} = \pi_B^T \land \pi_B^{T'} \). The discretization process consists of replacing each value that falls in the block \( Q_i \) of \( \pi_B^T \) by \( i \) for \( 0 \leq i \leq \ell \).
Active attribute domains partitions induce partitions of the set of objects. Namely, the partition of the set of objects $S$ that corresponds to a partition $\pi$ of $\text{adom}(B)$, where $B$ is a numerical attribute is denoted by $\pi_*$. A block of $\pi_*$ consists of all objects whose $B$-component belong to the same block of $\pi$. Note that when $\pi = \alpha_{\text{adom}(B)}$, then $\pi_* = \kappa_B$.

Suppose now that we have a numerical attribute $B$ and a categorical attribute $A$ that represents the class of each object $o_1, \ldots, o_\ell$, where

$$o_1[B] \leq o_2[B] \leq \ldots \leq o_n[B].$$
Define the partition $\pi_{B,A}$ of $\text{dom}(B)$ whose blocks consist of maximal subsequences $o_i[B], \ldots, o_{\ell}[B]$ of $o_1[B] \leq o_2[B] \leq \ldots \leq o_n[B]$ such that every object $o_j$ belongs to the same block of the partition $\pi^A$.

**Example**

Let $\{o_1, \ldots, o_9\}$ be a collection of nine objects sorted on the attribute $B$:

<table>
<thead>
<tr>
<th></th>
<th>$B$</th>
<th></th>
<th>$A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$o_1$</td>
<td>$\ldots$</td>
<td>95.2</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$o_2$</td>
<td>$\ldots$</td>
<td>110.1</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$o_3$</td>
<td>$\ldots$</td>
<td>120.0</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$o_4$</td>
<td>$\ldots$</td>
<td>125.5</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$o_5$</td>
<td>$\ldots$</td>
<td>130.1</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$o_6$</td>
<td>$\ldots$</td>
<td>140.0</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$o_7$</td>
<td>$\ldots$</td>
<td>140.5</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$o_8$</td>
<td>$\ldots$</td>
<td>168.2</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$o_9$</td>
<td>$\ldots$</td>
<td>190.5</td>
<td>$\ldots$</td>
</tr>
</tbody>
</table>
Example cont’d

The partition $\pi_{B,A,\ast}$ is

$$\pi_{B,A,\ast} = \{ \{o_1\}, \{o_2\}, \{o_3, o_4\}, \{o_5, o_6\}, \{o_7, o_8, o_9\} \}.$$ 

The blocks of this partition correspond to the longest subsequences of the sequence $(o_1, \ldots, o_9)$ that consists of objects that belong to the same $A$-class.
Fayyad showed that to obtain the least value of Shannon’s conditional entropy $H_\beta(p_i^A|\pi^T_{B^*})$ the cutpoints $t$ of $T$ must be chosen among the boundary points of the partition $\pi_{B,A,*}$. This is a powerful result that limits drastically the number of cut points and improves the tractability of the discretization.
We proposed a generalization of Fayyad’s discretization techniques that relies on the metric on partitions defined by $\beta$-entropy. We have shown that with an appropriate choice of the parameters of the discretization process the resulting decision trees are smaller, have fewer leaves, and display higher of accuracy as verified by stratified cross-validation. We have shown the following statement.

**Theorem**

Let $S$ be a collection of objects, where the class of an object is determined by the attribute $A$ and let $\beta \in (1, 2]$. Let $T$ is a set of cutpoints such that the conditional entropy $H_{\beta}(\kappa_A|\pi_{B^*}^T)$ is minimal among the set of cutpoints with the same number of elements, then $T$ consists of boundary points of the partition $\pi_{B,A,*}$ of $\text{adom}(A)$. 
Based on Theorem 29 we developed and tested a new discretization algorithm on several machine learning data sets from UCI data sets that have numerical attributes. After discretizations performed with several values of $\beta$ (typically $\beta \in \{1.5, 1.8, 1.9, 2\}$) we built the decision trees on the discretized data sets using the WEKA J48 variant of C4.5. The discretization technique has a significant impact of the size and accuracy of the decision trees. The experimental results suggest that an appropriate choice of $\beta$ can reduce significantly the size and number of leaves of the decision trees, roughly maintaining the accuracy (measured by stratified 5-fold cross validation) or even increasing the accuracy.
The size, number of leaves and accuracy of the trees are described below, where trees built using the Fayyad-Irani discretization method of J48 are designated as “standard”.

<table>
<thead>
<tr>
<th>Database</th>
<th>Discretization method</th>
<th>Size</th>
<th>Number of leaves</th>
<th>Accuracy (stratified cross-validation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>heart-c</td>
<td>standard</td>
<td>51</td>
<td>30</td>
<td>79.20</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.5$</td>
<td>20</td>
<td>14</td>
<td>77.36</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.8$</td>
<td>28</td>
<td>18</td>
<td>77.36</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.9$</td>
<td>35</td>
<td>22</td>
<td>76.01</td>
</tr>
<tr>
<td></td>
<td>$\beta = 2.0$</td>
<td>54</td>
<td>32</td>
<td>76.01</td>
</tr>
<tr>
<td>glass</td>
<td>standard</td>
<td>57</td>
<td>30</td>
<td>57.28</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.5$</td>
<td>32</td>
<td>24</td>
<td>71.02</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.8$</td>
<td>56</td>
<td>50</td>
<td>77.10</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.9$</td>
<td>64</td>
<td>58</td>
<td>67.57</td>
</tr>
<tr>
<td></td>
<td>$\beta = 2.0$</td>
<td>92</td>
<td>82</td>
<td>66.35</td>
</tr>
<tr>
<td>ionosphere</td>
<td>standard</td>
<td>35</td>
<td>18</td>
<td>90.88</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.5$</td>
<td>15</td>
<td>8</td>
<td>95.44</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.8$</td>
<td>19</td>
<td>12</td>
<td>88.31</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.9$</td>
<td>15</td>
<td>10</td>
<td>90.02</td>
</tr>
<tr>
<td></td>
<td>$\beta = 2.0$</td>
<td>15</td>
<td>10</td>
<td>90.02</td>
</tr>
<tr>
<td>iris</td>
<td>standard</td>
<td>9</td>
<td>5</td>
<td>95.33</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.5$</td>
<td>7</td>
<td>5</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.8$</td>
<td>7</td>
<td>5</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.9$</td>
<td>7</td>
<td>5</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>$\beta = 2.0$</td>
<td>7</td>
<td>5</td>
<td>96</td>
</tr>
<tr>
<td>diabetes</td>
<td>standard</td>
<td>43</td>
<td>22</td>
<td>74.08</td>
</tr>
<tr>
<td></td>
<td>$\beta = 1.8$</td>
<td>5</td>
<td>3</td>
<td>75.78</td>
</tr>
</tbody>
</table>
The notion of generalized entropy affords more flexibility in machine learning algorithm design. Decision trees, discretization techniques, and feature selection can benefit from using a generalized form of entropy and the metrics associated with it.
A research problem that we propose is the computation of consensus partitions using entropic metrics. When several partitions \( \pi_1, \ldots, \pi_n \) exist on a set \( S \) finding a consensus partition \( \pi \) aims to summarize these partitions. A general approach proposed is to consider a metric \( d \) on \( \text{PART}(S) \) and to seek a partition \( \pi \) on \( S \) such that the sum \( \sum_{i=1}^{n} d(\pi, \pi_i) \) is minimal. Entropic distances on \( \text{PART}(S) \) offer a natural background for solving this kind of problem.

Finding a consensus partition arises, for example, when consider several classifications that are regarded as approximations of a true classification that we need to recover, or when \( n \) partitions \( \pi_t, \pi_{t+1}, \ldots, \pi_{t+n-1} \) result from measurements at times \( t, t+1, \ldots, t+n-1 \) and one seeks a smoothing consensus of these partitions.
The proposed approach entails the construction a chain of partitions $\alpha_S = \sigma_0 \leq \sigma_1 \leq \cdots$, using an approach similar to hierarchical clustering and to seek $\sigma$ in this chain such that the sum $\sum_{i=1}^{n} d_\beta(\sigma, \pi_i)$ is minimal.