A Generalization of Conditional Entropy

Dan A. Simovici — Szymon Jaroszewicz

Univ. of Massachusetts Boston
Dept. of Computer Science
Boston, Massachusetts 02125 USA

{dsim,sj}@cs.umb.edu

ABSTRACT. *We introduce an extension of the notion of Shannon conditional entropy to a more general form of conditional entropy that captures both the conditional Shannon entropy and a similar notion related to the Gini index. The proposed family of conditional entropies generates a collection of metrics over the set of partitions of finite sets, which can be used to construct decision trees. Experimental results suggest that by varying the parameter that defines the entropy it is possible to obtain smaller decision trees for certain databases without sacrificing accurracy.*

RÉSUMÉ. *Nous présentons une extension de la notion de l'entropie conditionnelle de Shannon à une forme plus générale d'entropie conditionnelle qui formalise l'entropie conditionnelle de Shannon et une notion semblable liée à l'index de Gini. La famille proposée des entropies conditionnelles produit d'une collection de métriques sur l'ensemble de partitions des ensembles finis, qui peuvent être employées pour construire des arbres de décision. Les résultats expérimentaux suggèrent qu'en changeant le paramètre qui définit l'entropie il soit possible d'obtenir de plus petits arbres de décision pour certaines bases de données sans sacrifier l'exactitude de la classification.*

KEYWORDS: *Shannon entropy, Gini index, generalized conditional entropy, metric, partition, decision tree*

MOTS-CLÉS : *entropie de Shannon, index de Gini, entropie conditionnelle generalisée, métrique, partition, arbre de décision*

1. Introduction

Traditionally, the notion of Shannon entropy is introduced for a random variable distribution

$$X : \begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{pmatrix}$$

as $\mathcal{H}(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$. A partition $\pi = \{B_1, \ldots, B_m\}$ on a finite, nonempty set A generates naturally a random variable

$$X_\pi : \begin{pmatrix} B_1 & \cdots & B_m \\ \frac{|B_1|}{|S|} & \cdots & \frac{|B_m|}{|S|} \end{pmatrix}$$

We define the Shannon entropy of π as the Shannon entropy of X_π.

In [SIM 02] we introduced an axiomatization of a general notion of entropy for partitions of finite sets. Our system of axioms shows the common nature of Shannon entropy and of other measures of distribution concentration such that the Gini index. The goal of this paper is to introduce a metric on $\mathsf{PART}(A)$ starting from generalized conditional entropy of partitions. We show that these metrics generate selection criteria for splitting attributes in the construction of decision trees that result in smaller trees without any appreciable loss in accuracy.

Let $\mathsf{PART}(A)$ be the set of partitions of the nonempty set A. The class of all partitions of finite sets is denoted by PART. The one-block partition of A is denoted by ω_A, while the partition $\{\{a\} \mid a \in A\}$ is denoted by ι_A. If $\pi, \pi' \in \mathsf{PART}(A)$, then $\pi \leq \pi'$ if every block of π is included in a block of π'. Clearly, for every $\pi \in \mathsf{PART}(A)$ we have $\iota_A \leq \pi \leq \omega_A$. The partial ordered set $(\mathsf{PART}(A), \leq)$ is a lattice (see, for example a very lucid study of this lattice in [LER 81]). If $\sigma, \sigma' \in \mathsf{PART}(A)$, then σ' covers σ if $\sigma \leq \sigma'$ and there is no partition $\sigma_1 \in \mathsf{PART}(A)$ such that $\sigma \leq \sigma_1 \leq \sigma'$. This is denoted by $\sigma \prec \sigma'$. It is easy to see that $\sigma \prec \sigma'$ if and only if σ' can be obtained from σ by fusing two of its blocks into a block of σ'. The infimum of two partitions $\pi, \sigma \in \mathsf{PART}(A)$ will be denoted by $\pi \wedge \sigma$.

Partitions play a central role in classifications. Indeed, if a set of tuples T is described by attributes $a_1, \ldots, a_n$, then each set of attribute K defines a partition $\pi(K)$ of T, where two tuples belong to the same block of $\pi(K)$ if they have equal projections on K. Note that $H \subseteq K$, then $\pi(K) \leq \pi(H)$ for any attribute sets H and K.

If A, B are two disjoint and nonempty sets, $\pi \in \mathsf{PART}(A)$, $\sigma \in \mathsf{PART}(B)$, where $\pi = \{A_1, \ldots, A_m\}$, $\sigma = \{B_1, \ldots, B_n\}$, then the partition $\pi + \sigma$ is the partition of $A \cup B$ given by $\pi + \sigma = \{A_1, \ldots, A_m, B_1, \ldots, B_n\}$. Whenever the "$+$" operation is defined, then it is easily seen to be associative. In other words, if A, B, C are pairwise disjoint and nonempty sets, and $\pi \in \mathsf{PART}(A)$, $\sigma \in \mathsf{PART}(B)$, $\tau \in \mathsf{PART}(C)$, then $\pi + (\sigma + \tau) = (\pi + \sigma) + \tau$. Observe that if A, B are disjoint, then $\iota_A + \iota_B = \iota_{A \cup B}$. Also, $\omega_A + \omega_B$ is the partition $\{A, B\}$ of the set $A \cup B$.

Let $\pi = \{A_1, \ldots, A_m\} \in \mathsf{PART}(A)$ and $\sigma = \{B_1, \ldots, B_n\} \in \mathsf{PART}(B)$. The partition $\{A_i \times B_j \mid 1 \leq i \leq m, 1 \leq j \leq n\}$ of $A \times B$ is denoted by $\pi \times \sigma$. Note that $\iota_A \times \iota_B = \iota_{A \times B}$ and $\omega_A \times \omega_B = \omega_{A \times B}$.

The axiomatization introduced in [SIM 02] consists of four axioms satisfied by several types of entropy-like characteristics of partitions.

Definition 1.1 Let $\beta \in \mathbb{R}$, $\beta > 0$, and let $\Phi : \mathbb{R}^2_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ be a continuous function such that $\Phi(x, y) = \Phi(y, x)$, $\Phi(x, 0) = x$ for $x, y \in \mathbb{R}_{\geq 0}$.

A (Φ, β)*-system of axioms for a partition entropy* $\mathcal{H} : \mathsf{PART}(A) \longrightarrow \mathbb{R}_{\geq 0}$ consists of the following axioms:

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

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

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

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

(P4) If $\pi \in \mathsf{PART}(A)$ and $\sigma \in \mathsf{PART}(B)$, then $\mathcal{H}(\pi \times \sigma) = \Phi(\mathcal{H}(\pi), \mathcal{H}(\sigma))$.

□

Observe that we postulate that $\mathcal{H}(\pi) \geq 0$ for any partition π since the range of every function $\mathcal{H}$ is $\mathbb{R}_{\geq 0}$.

For a choice of β these axioms determine an entropy function $\mathcal{H}_\beta$ up to a constant factor. The same choice also determines the function Φ. The entropies defined for $\beta \neq 1$ were named *non-Shannon entropies*. In this case, for a partition $\pi = \{A_1, \ldots, A_n\} \in \mathsf{PART}(A)$ we have: $\mathcal{H}_\beta(\pi) = k\left(1 - \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta}\right)$, where k is a constant that satisfies the inequality $k(\beta - 1) > 0$. Thus, for $\beta > 1$ we have

$$\mathcal{H}_\beta(\pi) = c\left(1 - \sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta}\right),$$

and for $\beta < 1$ we have

$$\mathcal{H}_\beta(\pi) = c\left(\sum_{j=1}^{n} \left(\frac{|A_j|}{|A|}\right)^{\beta} - 1\right),$$

for some positive constant c, where $c = k$ if $\beta > 1$, and $c = -k$ when $\beta < 1$. In either case, we have $\Phi(x, y) = x + y - \frac{1}{k}xy$ for $x, y \in \mathbb{R}_{\geq 0}$.

The case $\beta = 1$ yields the Shannon entropy, that is

$$\mathcal{H}_1(\pi) = -c \sum_{j=1}^{n} \frac{|A_j|}{|A|} \log_2 \frac{|A_j|}{|A|}.$$

Also, if $\beta = 1$, then $\Phi(x, y) = x + y$ for $x, y \in \mathbb{R}_{\geq 0}$.

2. Metrics on Partitions Induced by Generalized Entropies

The generalized entropies previously introduced generate corresponding generalized conditional entropies. Let $\pi \in \mathsf{PART}(A)$ and let $C \subseteq A$. Denote by π_C the "trace" of π on C given by $\pi_C = \{B \cap C | B \in \pi \text{ such that } B \cap C \neq \emptyset\}$. Clearly, $\pi_C \in \mathsf{PART}(C)$; also, if C is a block of π, then $\pi_C = \omega_C$.

Definition 2.1 The *conditional entropy* defined by the (Φ, β)-entropy $\mathcal{H}$ is the function $\mathcal{H}_\beta : \mathsf{PART}^2 \longrightarrow \mathbb{R}_{\geq 0}$ given by: $\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \frac{|C_j|}{|A|} \cdot \mathcal{H}_\beta(\pi_{C_j})$, where $\pi, \sigma \in \mathsf{PART}(A)$ and $\sigma = \{C_1, \ldots, C_n\}$. □

Observe that $\mathcal{H}_\beta(\pi|\omega_A) = \mathcal{H}_\beta(\pi)$.

A direct consequence of the Axioms is that $\mathcal{H}(\omega_A) = 0$ for any set A (Lemma II.2 from [SIM 02]). The following reciprocal result also holds:

Lemma 2.2 *Let A be a finite set and let $\pi \in \mathsf{PART}(A)$ such that $\mathcal{H}(\pi) = 0$. Then, $\pi = \omega_A$.*

Proof. Suppose that $\mathcal{H}_\beta(\pi) = 0$ but $\pi < \omega_A$. Then, there exists a block C of π such that $\emptyset \subset C \subset A$. If $\theta = \{C, A - C\}$, then clearly we have $\pi \leq \theta$, so $0 \leq \mathcal{H}_\beta(\theta) \leq \mathcal{H}_\beta(\pi)$, which implies $\mathcal{H}_\beta(\theta) = 0$. If $\beta > 1$, then

$$\mathcal{H}_\beta(\theta) = c \left(1 - \left(\frac{|C|}{|A|} \right)^\beta - \left(\frac{|A - C|}{|A|} \right)^\beta \right) = 0.$$

The concavity of the function $f(x) = x^\beta + (1 - x)^\beta$ on $[0, 1]$ (when $\beta > 1$) implies either $C = A$ or $C = \emptyset$, which is a contradiction. Thus, $\pi = \omega_A$. A similar argument works for the other cases. ▮

Theorem 2.3 *Let A be a finite set and let $\pi, \sigma \in \mathsf{PART}(A)$. We have $\mathcal{H}_\beta(\pi|\sigma) = 0$ if and only if $\sigma \leq \pi$.*

Proof. Suppose that $\sigma = \{C_1, \ldots, C_n\}$. If $\sigma \leq \pi$, then $\pi_{C_j} = \omega_{C_j}$ for $1 \leq j \leq n$, so $\mathcal{H}_\beta(\pi|\sigma) = 0$. Conversely, suppose that

$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \frac{|C_j|}{|A|} \cdot \mathcal{H}_\beta(\pi_{C_j}) = 0.$$

This implies $\mathcal{H}_\beta(\pi_{C_j}) = 0$ for $1 \leq j \leq n$, so $\pi_{C_j} = \omega_{C_j}$ for $1 \leq j \leq n$ by Lemma 2.2. This means that every block C_j of σ is included in a block of π, which implies $\sigma \leq \pi$. ▮

Note that the partition $\pi \wedge \sigma$ whose blocks consist of nonempty intersections π and σ can be written as $\pi \wedge \sigma = \pi_{C_1} + \cdots + \pi_{C_n} = \sigma_{B_1} + \cdots + \sigma_{B_m}$. Therefore, by Corollary II.7 of [SIM 02], we have: $\mathcal{H}_\beta(\pi \wedge \sigma) = \sum_{j=1}^n \left(\frac{|C_j|}{|A|}\right)^\beta \mathcal{H}_\beta(\pi_{C_j}) + \mathcal{H}_\beta(\sigma)$.

For those entropies with $\beta > 1$ we have

$$\mathcal{H}_\beta(\pi \wedge \sigma) \leq \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma), \tag{1}$$

while for those having $\beta < 1$, the reverse inequality holds. In the case of Shannon entropy, $\beta = 1$ and

$$\begin{aligned} \mathcal{H}_1(\pi \wedge \sigma) &= \mathcal{H}_1(\pi|\sigma) + \mathcal{H}_1(\sigma) \qquad (2) \\ &= \mathcal{H}_1(\sigma|\pi) + \mathcal{H}_1(\pi). \end{aligned}$$

Lemma 2.4 *Let $a, b \in [0,1]$ such that $a + b = 1$. Then, for $\beta > 1$ we have:*

$$\sum_{i=1}^n (ax_i + by_i)^\beta \leq a \sum_{i=1}^n x_i^\beta + b \sum_{i=1}^n y_i^\beta,$$

for every $x_1, \ldots, x_n, y_1, \ldots, y_n \in [0,1]$. For $\beta < 1$, the reverse inequality holds.

Proof. The statement follows immediately from concavity of the function $f(x) = x^\beta$ for $\beta > 1$ on the interval $[0,1]$. ▮

Theorems 2.5 and 2.8 extend well-known monotonicity properties of Shannon entropy.

Theorem 2.5 *If π, σ, σ' are partitions of the finite set A such that $\sigma \leq \sigma'$, then $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi|\sigma')$ for $\beta > 0$.*

Proof. To prove this statement it suffices to consider only the case when $\sigma \prec \sigma'$. Suppose initially that $\beta > 1$.

Let $\sigma, \sigma' \in \mathsf{PART}(A)$ such that $\sigma \prec \sigma'$. Suppose that D, E are blocks of σ such that $C = D \cup E$, where C is a block of σ'; the partition π is $\{B_1, \ldots, B_n\}$.

Define $x_i = \frac{|B_i \cap D|}{|D|}$ and $y_i = \frac{|B_i \cap E|}{|E|}$ for $1 \leq i \leq n$. If we choose $a = \frac{|D|}{|C|}$ and $b = \frac{|E|}{|C|}$, then

$$|C| \sum_{i=1}^n \frac{|B_i \cap C|^\beta}{|C|^\beta} \leq |D| \sum_{i=1}^n \frac{|B_i \cap D|^\beta}{|D|^\beta} + |E| \sum_{i=1}^n \frac{|B_i \cap E|^\beta}{|E|^\beta},$$

by Lemma 2.4. Consequently, we can write:

$$\begin{aligned}
\mathcal{H}_\beta(\pi|\sigma) &= \cdots + \frac{|D|}{|A|}\mathcal{H}_\beta(\pi_D) + \frac{|E|}{|A|}\mathcal{H}_\beta(\pi_E) + \cdots \\
&= \cdots + \frac{|D|}{|A|}\left(1 - \sum_{i=1}^{n} \frac{|B_i \cap D|^\beta}{|D|^\beta}\right) + \frac{|E|}{|A|}\left(1 - \sum_{i=1}^{n} \frac{|B_i \cap E|^\beta}{|E|^\beta}\right) + \cdots \\
&\leq \cdots + \frac{|C|}{|A|}\left(1 - \sum_{i=1}^{n} \frac{|B_i \cap C|^\beta}{|C|^\beta}\right) + \cdots = \mathcal{H}_\beta(\pi|\sigma').
\end{aligned}$$

For $\beta < 1$ we have

$$|C|\sum_{i=1}^{n} \frac{|B_i \cap C|^\beta}{|C|^\beta} \geq |D|\sum_{i=1}^{n} \frac{|B_i \cap D|^\beta}{|D|^\beta} + |E|\sum_{i=1}^{n} \frac{|B_i \cap E|^\beta}{|E|^\beta},$$

by the second part of Lemma 2.4. Thus,

$$\begin{aligned}
\mathcal{H}_\beta(\pi|\sigma) &= \cdots + \frac{|D|}{|A|}\mathcal{H}_\beta(\pi_D) + \frac{|E|}{|A|}\mathcal{H}_\beta(\pi_E) + \cdots \\
&= \cdots + \frac{|D|}{|A|}\left(\sum_{i=1}^{n} \frac{|B_i \cap D|^\beta}{|D|^\beta} - 1\right) + \frac{|E|}{|A|}\left(\sum_{i=1}^{n} \frac{|B_i \cap E|^\beta}{|E|^\beta} - 1\right) + \cdots \\
&\leq \cdots + \frac{|C|}{|A|}\left(\sum_{i=1}^{n} \frac{|B_i \cap C|^\beta}{|C|^\beta} - 1\right) + \cdots = \mathcal{H}_\beta(\pi|\sigma').
\end{aligned}$$

For $\beta = 1$ the inequality is a well-known property of Shannon entropy. ∎

Corollary 2.6 *For every $\pi, \sigma \in \mathsf{PART}(A)$ and $\beta > 0$, we have $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi)$.*

Proof. Since $\sigma \leq \omega_A$, by Theorem 2.5 we have $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi|\omega_A) = \mathcal{H}_\beta(\pi)$. ∎

Corollary 2.7 *Let A be a finite set. For $\beta \geq 1$ we have $\mathcal{H}_\beta(\pi \wedge \sigma) \leq \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma)$ for every $\pi, \sigma \in \mathsf{PART}(A)$.*

Proof. By Inequality (1) and by Corollary 2.6 we have

$$\mathcal{H}_\beta(\pi \wedge \sigma) \leq \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma) \leq \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma).$$

∎

Theorem 2.8 *If π, π', σ are partitions of the finite set A such that $\pi \leq \pi'$, then $\mathcal{H}_\beta(\pi|\sigma) \geq \mathcal{H}_\beta(\pi'|\sigma)$.*

Proof. Suppose that $\sigma = \{C_1, \ldots, C_n\}$. Then, it is clear that $\pi_{C_j} \leq \pi'_{C_j}$ for $1 \leq j \leq n$. Therefore, $\mathcal{H}_\beta(\pi_{C_j}) \geq \mathcal{H}_\beta(\pi'_{C_j})$ by Axiom **(P1)**, which implies immediately the desired inequality. ▮

Lemma 2.9 *Let A be a nonempty set and let $\{A', A''\}$ be a two-block partition of A. If $\pi \in \mathsf{PART}(A)$, $\sigma' \in \mathsf{PART}(A')$, and $\sigma'' \in \mathsf{PART}(A'')$, then*

$$\mathcal{H}_\beta(\pi|\sigma' + \sigma'') = \frac{|A'|}{|A|}\mathcal{H}_\beta(\pi'|\sigma') + \frac{|A''|}{|A|}\mathcal{H}_\beta(\pi''|\sigma''),$$

where $\pi' = \pi_{A'}$ and $\pi'' = \pi_{A''}$.

Proof. Note that $\sigma' + \sigma''$ is a partition of A. The lemma follows immediately from the definition of conditional entropy. ▮

Theorem 2.10 *Let A be a nonempty set and let $\{A_1, \ldots, A_\ell\}$ be a partition of A. If $\pi \in \mathsf{PART}(A)$, $\sigma_k \in \mathsf{PART}(A_k)$ for $1 \leq k \leq \ell$, then*

$$\mathcal{H}_\beta(\pi|\sigma_1 + \cdots + \sigma_\ell) = \sum_{k=1}^{\ell} \frac{|A_k|}{|A|}\mathcal{H}_\beta(\pi_k|\sigma_k)$$

where $\pi_k = \pi_{A_k}$ for $1 \leq k \leq \ell$.

Proof. The result follows immediately from Lemma 2.9 due to the associativity of the partial operation "$+$". ▮

Theorem 2.11 *If $\beta > 1$, then for every three partitions π, σ, τ of a finite set A we have*

$$\mathcal{H}_\beta(\pi|\sigma \wedge \tau) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi \wedge \sigma|\tau).$$

If $\beta < 1$ we have the reverse inequality, and for $\beta = 1$ we have the equality

$$\mathcal{H}_\beta(\pi|\sigma \wedge \tau) + \mathcal{H}_\beta(\sigma|\tau) = \mathcal{H}_\beta(\pi \wedge \sigma|\tau).$$

Proof. Suppose that $\pi = \{B_1, \ldots, B_m\}$, $\sigma = \{C_1, \ldots, C_m\}$, and $\tau = \{D_1, \ldots, D_\ell\}$. We noted already that $\sigma \wedge \tau = \sigma_{D_1} + \cdots + \sigma_{D_\ell} = \tau_{C_1} + \cdots + \tau_{C_n}$. Consequently, by Theorem 2.10 we have $\mathcal{H}_\beta(\sigma \wedge \pi) = \sum_{k=1}^{\ell} \frac{|D_k|}{|A|}\mathcal{H}_\beta(\pi_{D_k}|\sigma_{D_k})$. Also, we have $\mathcal{H}_\beta(\sigma|\tau) = \sum_{k=1}^{\ell} \frac{|D_k|}{|A|}\mathcal{H}_\beta(\sigma_{D_k})$.

If $\beta > 1$ we saw that $\mathcal{H}_\beta(\pi_{D_k} \wedge \sigma_{D_k}) \leq \mathcal{H}_\beta(\pi_{D_k}|\sigma_{D_k}) + \mathcal{H}_\beta(\sigma_{D_k})$, for every k, $1 \leq k \leq \ell$, which implies

$$\begin{aligned}
\mathcal{H}_\beta(\sigma \wedge \pi) + \mathcal{H}_\beta(\sigma|\tau) &\geq \sum_{k=1}^{\ell} \frac{|D_k|}{|A|} \mathcal{H}_\beta(\pi_{D_k} \wedge \sigma_{D_k}) \\
&= \sum_{k=1}^{\ell} \frac{|D_k|}{|A|} \mathcal{H}_\beta((\pi \wedge \sigma)_{D_k}) \\
&= \mathcal{H}_\beta(\pi \wedge \sigma|\tau).
\end{aligned}$$

Using a similar argument we obtain the second inequality of the theorem. The equality for the Shannon case was obtained in [MÁN 91]. ▮

Corollary 2.12 *Let A be a finite set. For $\beta \geq 1$ and for $\pi, \sigma, \tau \in \mathsf{PART}(A)$ we have the inequality: $\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi|\tau)$.*

Proof. Note that by Theorem 2.5 we have: $\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi|\sigma \wedge \tau) + \mathcal{H}_\beta(\sigma|\tau)$. Therefore, for $\beta \geq 1$, by Theorems 2.11 and 2.8 we obtain $\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\tau) \geq \mathcal{H}_\beta(\pi \wedge \sigma|\tau) \geq \mathcal{H}_\beta(\pi|\tau)$. ▮

Definition 2.13 Let $\beta > 1$. The mapping $d_\beta : \mathsf{PART}(A)^2 \longrightarrow \mathbb{R}_{\geq 0}$ is defined by $d_\beta(\pi, \sigma) = \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi)$ for $\pi, \sigma \in \mathsf{PART}(A)$. □

The following result generalizes a result of López de Mántaras:

Corollary 2.14 *d_β is a metric on $\mathsf{PART}(A)$.*

Proof. If $d_\beta(\pi, \sigma) = 0$, then $\mathcal{H}_\beta(\pi|\sigma) = \mathcal{H}_\beta(\sigma|\pi) = 0$. Therefore, by Theorem 2.3 we have $\sigma \leq \pi$ and $\pi \leq \sigma$, so $\pi = \sigma$. The symmetry of d_β is immediate. The triangular property is a direct consequence of Corollary 2.12. ▮

In [MÁN 91] it is shown that the mapping $e_1 : \mathsf{PART}(A)^2 \longrightarrow \mathbb{R}_{\geq 0}$ that corresponds to Shannon entropy, defined by

$$e_1(\pi, \sigma) = \frac{d_1(\pi, \sigma)}{\mathcal{H}_1(\pi \wedge \sigma)}$$

for $\pi, \sigma \in \mathsf{PART}(A)$ is also a metric on $\mathsf{PART}(A)$. This result is extended next.

Theorem 2.15 *Let A be a finite, non-empty set. For $\beta \geq 1$, the mapping $e_\beta : \mathsf{PART}(A)^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by*

$$e_\beta(\pi, \sigma) = \frac{2 d_\beta(\pi, \sigma)}{d_\beta(\pi, \sigma) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma)}$$

for $\pi, \sigma \in \mathsf{PART}(A)$ is a metric on $\mathsf{PART}(A)$ such that $0 \leq e_\beta(\pi, \sigma) \leq 1$.

Proof. It easy to see that $0 \leq e_\beta(\pi,\sigma) \leq 1$ since, by Corollary 2.6, $\mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma) \geq \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) = d_\beta(\pi,\sigma)$. We need to show only that the triangular inequality is satisfied by e_β for $\beta > 1$. We can write:

$$\begin{array}{l} e_\beta(\pi,\sigma) + e_\beta(\sigma,\tau) = \\ \frac{\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)}{\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)+\mathcal{H}_\beta(\pi)+\mathcal{H}_\beta(\sigma)} + \frac{\mathcal{H}_\beta(\sigma|\tau)+\mathcal{H}_\beta(\tau|\sigma)}{\mathcal{H}_\beta(\sigma|\tau)+\mathcal{H}_\beta(\tau|\sigma)+\mathcal{H}_\beta(\sigma)+\mathcal{H}_\beta(\tau)}. \end{array}$$

Note that

$$\begin{array}{l} \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma) \leq \\ \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau|\sigma) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\tau) \end{array}$$

because $\mathcal{H}_\beta(\sigma) \leq \mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau)$ by Inequality (1) and Axiom **(P1)**. Similarly,

$$\begin{array}{l} \mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau|\sigma) + \mathcal{H}_\beta(\sigma) + \mathcal{H}_\beta(\tau) \leq \\ \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\sigma|\tau) + \mathcal{H}_\beta(\tau|\sigma) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\tau) \end{array}$$

because $\mathcal{H}_\beta(\sigma) \leq \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi)$. This yields the inequality:

$$\begin{array}{l} e_\beta(\pi,\sigma) + e_\beta(\sigma,\tau) \geq \\ \frac{\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)+\mathcal{H}_\beta(\sigma|\tau)+\mathcal{H}_\beta(\tau|\sigma)}{\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)+\mathcal{H}_\beta(\sigma|\tau)+\mathcal{H}_\beta(\tau|\sigma)+\mathcal{H}_\beta(\pi)+\mathcal{H}_\beta(\tau)} = \\ \frac{1}{1+\frac{\mathcal{H}_\beta(\pi)+\mathcal{H}_\beta(\tau)}{\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)+\mathcal{H}_\beta(\sigma|\tau)+\mathcal{H}_\beta(\tau|\sigma)}} \geq \\ \frac{1}{1+\frac{\mathcal{H}_\beta(\pi)+\mathcal{H}_\beta(\tau)}{\mathcal{H}_\beta(\pi|\tau)+\mathcal{H}_\beta(\tau|\pi)}} = e_\beta(\pi,\tau). \end{array},$$

∎

For $\beta = 1$, $e_1(\pi,\sigma) = \frac{d_1(\pi,\sigma)}{\mathcal{H}_1(\pi\wedge\sigma)}$, due to equality (2), which coincides with the expression obtained in [MÁN 91] for the normalized distance.

3. Generalized Gain as a Selection Criterion for Splitting Attributes in Decision Trees

The standard selection criterion for splitting attributes is the information gain used by Quinlan [QUI 93] in the classical C4.5 algorithm. We show that choosing the splitting attribute A based on the least value of $d_\beta(\pi(A),\pi)$, where π is the partition of the training set that corresponds to the target attribute of the classification generates smaller trees with comparable degrees of accuracy.

Let $\pi,\sigma \in \mathsf{PART}(A)$. The *$\beta$-gain* of σ relative to π is the expression $G_\beta(\pi,\sigma) = \mathcal{H}_\beta(\pi) - \mathcal{H}_\beta(\pi|\sigma)$. The gain ratio is given by $R_\beta(\pi,\sigma) = \frac{G_\beta(\pi,\sigma)}{\mathcal{H}_\beta(\sigma)}$. For $\beta = 1$ we obtain Quinlan's gain defined through Shannon's entropy.

The next theorem establishes a monotonicity property of the distance d_β that shows that d_β does not favor attributes with large domains, an issue that is important for building decision trees.

Theorem 3.1 *Let A be a finite set and let $\pi, \pi', \sigma \in \mathsf{PART}(A)$ be such that π' is covered by π. In other words, $\pi = \{B_1, \ldots, B_m\}$ and $\pi' = \{B_1, \ldots, B'_m, B''_m\}$, where $B_m = B'_m \cup B''_m$. Suppose also that there exists a block C of σ such that $B_m \subseteq C$. Then, if $\beta \geq 1$, we have $d_\beta(\pi, \sigma) \leq d_\beta(\pi', \sigma)$ and $e_\beta(\pi, \sigma) \leq e_\beta(\pi', \sigma)$.*

Proof. For the case of Shannon's entropy, $\beta = 1$, the inequalities were proven in [MÁN 91]. Therefore, we can assume that $\beta > 1$.

We claim that under the hypothesis of the theorem we have $\mathcal{H}_\beta(\sigma|\pi) = \mathcal{H}(\sigma|\pi')$. Note that $\sigma_{B_m} = \omega_{B_m}$, $\sigma_{B'_m} = \omega_{B'_m}$, and $\sigma_{B''_m} = \omega_{B''_m}$, since $B'_m, B''_m \subseteq B_m \subseteq C$. Therefore, $\mathcal{H}(\sigma_{B_m}) = \mathcal{H}(\sigma_{B'_m}) = \mathcal{H}(\sigma_{B''_m}) = 0$, hence

$$\mathcal{H}_\beta(\sigma|\pi) = \sum_{i=1}^{m} \frac{|B_i|}{|A|}\mathcal{H}_\beta(\sigma_{B_i}) = \sum_{i=1}^{m-1} \frac{|B_i|}{|A|}\mathcal{H}_\beta(\sigma_{B_i}) = \mathcal{H}_\beta(\sigma|\pi').$$

Theorem 2.8 implies $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi'|\sigma)$, which gives the first inequality.

Note that the second equality of the theorem:

$$\begin{aligned}\mathcal{H}_\beta(\pi|\sigma) &= \frac{\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi)}{\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi) + \mathcal{H}_\beta(\sigma)} \\ &\leq \frac{\mathcal{H}_\beta(\pi'|\sigma) + \mathcal{H}_\beta(\sigma|\pi')}{\mathcal{H}_\beta(\pi'|\sigma) + \mathcal{H}_\beta(\sigma|\pi') + \mathcal{H}_\beta(\pi') + \mathcal{H}_\beta(\sigma)} = e_\beta(\pi', \sigma)\end{aligned}$$

is equivalent to

$$\frac{\mathcal{H}_\beta(\sigma) + \mathcal{H}_\beta(\pi)}{\mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi)} \geq \frac{\mathcal{H}_\beta(\sigma) + \mathcal{H}_\beta(\pi')}{\mathcal{H}_\beta(\pi'|\sigma) + \mathcal{H}_\beta(\sigma|\pi')}. \tag{3}$$

Applying the definition of conditional entropy we can write:

$$\mathcal{H}_\beta(\pi|\sigma) - \mathcal{H}_\beta(\pi'|\sigma) = \frac{|C|}{|A|}\left[\frac{|B'_m|^\beta}{|C|^\beta} + \frac{|B''_m|^\beta}{|C|^\beta} - \frac{|B_m|^\beta}{|C|^\beta}\right]$$

and

$$\mathcal{H}_\beta(\pi) - \mathcal{H}_\beta(\pi') = \frac{|B'_m|^\beta + |B''_m|^\beta - |B_m|^\beta}{|A|^\beta},$$

which implies

$$\mathcal{H}_\beta(\pi|\sigma) - \mathcal{H}_\beta(\pi'|\sigma) = \left(\frac{|A|}{|C|}\right)^{\beta-1}\left[\mathcal{H}_\beta(\pi) - \mathcal{H}_\beta(\pi')\right]. \tag{4}$$

Thus, we obtain:

$$\mathcal{H}_\beta(\pi'|\sigma) - \mathcal{H}_\beta(\pi|\sigma) \geq \mathcal{H}_\beta(\pi') - \mathcal{H}_\beta(\pi). \tag{5}$$

Database	J48	$\beta = 1$	$\beta = 1.5$	$\beta = 2$	$\beta = 2.5$
audiology	78.76%	73.42%	73.86%	73.86%	71.64%
hepatitis	78.06%	83.22%	83.22%	83.87%	83.87%
primary-tumor	40.99%	43.34%	41.87%	43.34%	43.05%

Table 1. *Accuracy Results*

By denoting $a = \mathcal{H}_\beta(\sigma)$ and $b = \mathcal{H}_\beta(\sigma|\pi) = \mathcal{H}_\beta(\sigma|\pi')$, the Inequality (3) can be written as:

$$\frac{a + \mathcal{H}_\beta(\pi)}{\mathcal{H}_\beta(\pi|\sigma) + b} \geq \frac{a + \mathcal{H}_\beta(\pi')}{\mathcal{H}_\beta(\pi'|\sigma) + b},$$

Elementary transformations yield: $\mathcal{H}_\beta(\pi'|\sigma) - \mathcal{H}_\beta(\pi|\sigma) \geq \frac{b+\mathcal{H}_\beta(\pi|\sigma)}{a+\mathcal{H}_\beta(\pi)}(\mathcal{H}_\beta(\pi') - \mathcal{H}_\beta(\pi))$, which is implied by Inequality (5) because $\frac{b+\mathcal{H}_\beta(\pi|\sigma)}{a+\mathcal{H}_\beta(\pi)} \leq 1$. This proves the second inequality of the theorem. ▮

4. Experimental Results and Conclusions

The experiments have been conducted on 33 datasets from the UCI Machine Learning Repository. The J48 tree builder from the Weka package [WIT 00] was used, in its original form as well as modified to support generalized entropies for different values of the β parameter. Each experiment used 5-fold crossvalidation, average has been taken of the outcomes of the 5 runs and was performed with and without pruning.

The tree size and the number of leaves diminish for 20 of the 33 databases analysed and grow for the remaining 13. The best reduction in size was achieved for the primary-tumor database, where the size of the tree was reduced to 37% for $\beta = 2.5$ and the number of leaves was reduced to 38.8% compared to the standard J48 algorithm that makes use of the gain ratio. On another hand, the largest increase in size and number of leaves was recorded for the pima-diabetes database, where for $\beta = 1$, we has an increase to 260% in size and to 256% in the number of leaves, though such an increase occurs rarely among the 13 databases where increases occur.

In Figure 1 we show the comparative performance of the distance d_β approach compared to the standard gain ratio for the databases which yielded the best results (audiology, hepatitis, and primary-tumor), in the case of the prunned trees. The 100% level refers in each case to the gain-ratio algorithm. It is interesting to observe that the accurracy diminishes slightly (by % for audiology database) or improves slightly, as shown in table 1, thus confirming previous results [MÁN 91, BRE 98, MIN 89] that accuracy is not affected substantially by the method used for tree construction.

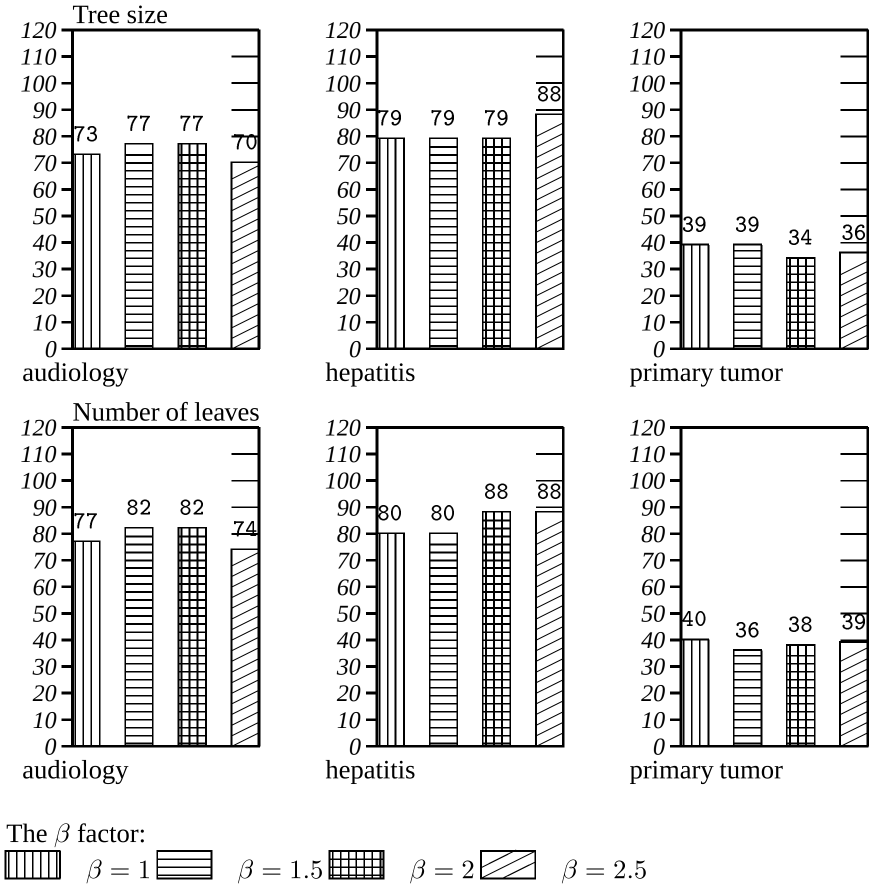

Figure 1. *Comparative Experimental Results*

5. References

[BRE 98] Breiman L., Friedman J. H., Olshen R. A., Stone C. J., *Classification and Regression Trees*, Chapman and Hall, Boca Raton, 1998.

[LER 81] Lerman I. C., *Classification et analyse ordinale des données*, Dunod, Paris, 1981.

[MÁN 91] de Mántaras R. L., “A Distance-Based Attribute Selection Measure for Decision Tree Induction”, *Machine Learning*, vol. 6, 1991, p. 81–92.

[MIN 89] Mingers J., “An Empirical Comparison of Selection Measures for Decision Tree Induction”, *Machine Learning*, vol. 3, 1989, p. 319–342.

[QUI 93] Quinlan J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

[SIM 02] Simovici D. A., Jaroszewicz S., “An Axiomatization of Partition Entropy”, *IEEE Transactions on Information Theory*, vol. 48, 2002, p. 2138–2142.

[WIT 00] Witten I. H., Frank E., *Data Mining*, Morgan-Kaufmann, San Francisco, 2000.