A New Metric Splitting Criterion for Decision Trees

Dan A. Simovici* Szymon Jaroszewicz†

Abstract: We examine a new approach to building decision tree by introducing a geometric splitting criterion, based on the properties of a family of metrics on the space of partitions of a finite set. This criterion can be adapted to the characteristics of the data sets and the needs of the users and yields decision trees that have smaller sizes and fewer leaves than the trees built with standard methods and have comparable or better accuracy.

Keywords: decision tree, generalized conditional entropy, metric, metric betweenness

1 Introduction

Decision trees constitute one of the most popular classification techniques in data mining and have been the subject of a large body of investigation. The typical construction algorithm for a decision tree starts with a training set of objects that is split recursively. The successive splits form a tree where the sets assigned to the leaves consist of objects that belong almost entirely to a single class. This allows new objects that belong to a test set to be classified into a specific class based on the path induced by the object in the decision tree which joins the root of the tree to a leaf.

Decision trees are useful classification algorithms, even though they may present problems related to overfitting and excessive data fragmentation that results in rather complex classification schemes.

A central problem in the construction of decision trees is the choice of the splitting attribute at each non-leaf node. We show that the usual splitting criterion (the information gain ratio, or the similar measure derived from the Gini index) are special cases of a more general approach. Furthermore, we propose a geometric criterion for choosing the splitting attributes that has the advantage of being adaptable to various data sets and user needs.

*University of Massachusetts at Boston, Dept. of Computer Science, Boston, Massachusetts 02125, e-mail: dsim@cs.umb.edu

†Faculty of Computer and Information Systems, Technical University of Szeczin, Poland, e-mail: sjaroszewicz@wi.ps.pl

2 Partition Entropies

A *metric* on a set S is a mapping $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ such that $d(s,t) = 0$ if and only if $s = t$, $d(s,t) = d(t,s)$ and $d(s,u) + d(u,t) \geq d(s,t)$ for every $s, t, u \in S$. The last inequality is known as the *triangular inequality*.

A *metric space* is a pair (S, d), where S is a set and d is a metric on S.

The *betweenness relation* defined by the metric space (S, d) is a ternary relation R on the set S defined by $(s, u, t) \in R$ if $d(s,u) + d(u,t) = d(s,t)$. We denote the fact that $(s, u, t) \in R$ by $[sut]$ and we say that u *is between* s *and* t.

We explore a natural link that exists between random variables and partitions of sets that allows the transfer of certain probabilistic and information-theoretical notions to partitions of sets.

A *partition* of a set S is a non-empty collection π of non-empty subsets of S, $\pi = \{B_i \mid i \in I\}$ such that for every $i, j \in I$, $i \neq j$ implies $B_i \cap B_j = \emptyset$ and $\bigcup_{i \in I} B_i = S$. We refer to the sets B_i as the *blocks* of π.

Let $\mathsf{PART}(S)$ be the set of partitions of a set S. The class of all partitions of finite sets is denoted by PART. The one-block partition of S is denoted by ω_S. The partition $\{\{s\} \mid s \in S\}$ is denoted by ι_S.

If $\pi, \pi' \in \mathsf{PART}(S)$, then $\pi \leq \pi'$ if every block of π is included in a block of π'. Clearly, for every $\pi \in \mathsf{PART}(S)$ we have $\iota_S \leq \pi \leq \omega_S$.

π' covers π if $\pi \leq \pi'$ and there is no partition $\theta \in \mathsf{PART}(S)$ such that $\pi < \theta < \pi'$. This fact is denoted by $\pi \prec \pi'$. It is known [Ler81] that $\pi \prec \pi'$ if and only if π' is obtained from π by fusing two blocks of this partition into a new block.

For every two partitions π, σ both $\inf\{\pi, \sigma\}$ and $\sup\{\pi, \sigma\}$ in the partial ordered set $(\mathsf{PART}(S), \leq)$ exist. Namely, if $\pi = \{B_i \mid i \in I\}$ and $\sigma = \{C_j \mid j \in J\}$, then $\inf\{\pi, \sigma\}$ is the partition:

$$\pi \wedge \sigma = \{B_i \cap C_j \mid B_i \cap C_j \neq \emptyset, i \in, j \in J\}.$$

The supremum $\pi \vee \sigma = \sup\{\pi, \sigma\}$ can be described using a bipartitite graph $\mathcal{G}$ having $\{B_i \mid i \in I\} \cup \{C_j \mid j \in J\}$ as set of vertices. An edge (B_i, C_j) exists only if $B_i \cap C_j \neq \emptyset$. If $\mathcal{C}$ is a connected component of $\mathcal{G}$ note that $\bigcup\{B \in \pi \mid B \in \mathcal{C}\} = \bigcup\{C \in \sigma \mid C \in \mathcal{C}\}$; we denote this set by $D_{\mathcal{C}}$. The family of sets $\{D_{\mathcal{C}} \mid \mathcal{C}$ is a connected component of $\mathcal{G}\}$ is a partition of the set S. It is easy to verify that this is exactly $\pi \vee \sigma$.

It is not difficult to show that $(\mathsf{PART}(S), \leq)$ is an upper semimodular lattice; in other words if π, σ are two distinct partitions such each covers $\pi \wedge \sigma$, then $\pi \vee \sigma$ covers both π and σ.

If S, T are two disjoint and nonempty sets, $\pi \in \mathsf{PART}(S)$, $\sigma \in \mathsf{PART}(T)$, where $\pi = \{A_1, \ldots, A_m\}$, $\sigma = \{B_1, \ldots, B_n\}$, then the partition $\pi + \sigma$ is the partition of $S \cup T$ 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 S, U, V are pairwise disjoint and nonempty sets, and $\pi \in \mathsf{PART}(S)$, $\sigma \in \mathsf{PART}(U)$, $\tau \in \mathsf{PART}(V)$, then $\pi + (\sigma + \tau) = (\pi + \sigma) + \tau$. Observe that if S, U are disjoint, then $\iota_S + \iota_U = \iota_{S \cup U}$. Also, $\omega_S + \omega_U$ is the partition $\{S, U\}$ of the set $S \cup U$.

If $\pi = \{B_1, \ldots, B_m\}$, $\sigma = \{C_1, \ldots, C_n\}$ are partitions of two arbitrary sets S, U, respectively, then we denote the partition $\{B_i \times C_j \mid 1 \leq i \leq m, 1 \leq j \leq n\}$ of $S \times U$ by $\pi \times \sigma$. Note that $\iota_S \times \iota_U = \iota_{S\times U}$ and $\omega_S \times \omega_U = \omega_{S\times U}$.

Let $\pi \in \mathsf{PART}(S)$ and let $C \subseteq S$. 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 $\pi_C = \omega_C$.

A subset T of S is *pure* relative to a partition $\pi \in \mathsf{PART}(S)$ if $\pi_T = \omega_T$. In other words, T is pure relative to a partition π if T is included in some block of π.

In [Dar70] the notion of β-entropy of a probability distribution $\mathbf{p} = (p_1, \ldots, p_n)$ was defined as:

$$\mathcal{H}_\beta(\mathbf{p}) = \frac{1}{2^{1-\beta} - 1} \left(\sum_{i=1}^{m} p_i^\beta - 1 \right),$$

where $p_1 + \cdots + p_n = 1$ and $p_i \geq 0$ for $1 \leq i \leq n$. In the same reference it was observed that Shannon's entropy $\mathcal{H}(\mathbf{p})$ can be obtained as $\lim_{\beta \to 1} \mathcal{H}_\beta(\pi)$.

In [SJ02] we offered a new interpretation of the notion of entropy for finite distributions as entropies of partitions of finite sets. Our approach takes advantage of the properties of the partial order of the lattice of partitions of a finite set and makes use of operations defined on partitions.

We defined the $\mathcal{H}_\beta$ entropy for $\beta \in \mathbb{R}$, $\beta > 0$ as a function $\mathcal{H}_\beta : \mathsf{PART}(S) \longrightarrow \mathbb{R}_{\geq 0}$ that satisfies conditions:

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

(P2) If S, T are two finite sets such that $|S| \leq |T|$, then $\mathcal{H}(\iota_S) \leq \mathcal{H}(\iota_T)$.

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

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

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

where $\Phi : \mathbb{R}^2_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ be a continuous function such that $\Phi(x, y) = \Phi(y, x)$, $\Phi(x, 0) = x$ for $x, y \in \mathbb{R}_{\geq 0}$.

We have shown in [SJ02] that if $\pi = \{B_1, \ldots, B_n\} \in \mathsf{PART}(S)$, then

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

In the special case, when $\beta \to 1$ we have:

$$\mathcal{H}_\beta(\pi) = -\sum_{i=1}^{m} \frac{|B_i|}{|S|} \cdot \log_2 \frac{|B_i|}{|S|}.$$

The axiomatization also implies a specific form of the function Φ. Namely, if $\beta \neq 1$ it follows that $\Phi(x, y) = x + y + (2^{1-\beta} - 1)xy$. In the case of Shannon entropy, obtained using $\beta = 1$, we have $\Phi(x, y) = x + y$ for $x, y \in \mathbb{R}_{\geq 0}$.

Note that if $|S| = 1$, then $\mathsf{PART}(S)$ consists of a unique partition ($\omega_S = \iota_S$) and $\mathcal{H}_\beta(\omega_S) = 0$. Moreover, for an arbitrary finite set S we have $\mathcal{H}_\beta(\pi) = 0$ if and only if $\pi = \omega_S$. Indeed, let U, V be two finite disjoint sets that have the same cardinality. Axiom **(P3)** implies:

$$\mathcal{H}_\beta(\omega_U + \omega_V) = \left(\frac{1}{2}\right)^\beta (\mathcal{H}(\omega_U) + \mathcal{H}_\beta(\omega_V)) + \mathcal{H}_\beta(\{U, V\}).$$

Since $\omega_U + \omega_V = \{U, V\}$ it follows that $\mathcal{H}_\beta(\omega_U) = \mathcal{H}_\beta(\omega_V) = 0$.

Conversely, suppose that $\mathcal{H}_\beta(\pi) = 0$. If $\pi < \omega_S$ there exists a block B of of π such that $\emptyset \subset B \subset S$. Let θ be the partition $\theta = \{B, S - B\}$. It is clear that $\pi \leq \theta$, so $0 \leq \mathcal{H}_\beta(\theta) \leq \mathcal{H}_\beta(\pi)$ which implies $\mathcal{H}_\beta(\theta) = 0$. This, in turn yields:

$$\left(\frac{|B|}{|S|}\right)^\beta \left(\frac{|S - B|}{|S|}\right)^\beta - 1 = 0$$

Since the function $f(x) = x^\beta + (1 - x)^\beta$ is concave for $\beta > 1$ and convex for $\beta < 1$ on the interval $[0, 1]$, the above equality is possible only if $B = S$ or if $B = \emptyset$, which is a contradiction. Thus, $\pi = \omega_S$.

These facts suggest that for a subset T of S the number $\mathcal{H}_\beta(\pi_T)$ can be used as a measure of the purity of the set T with respect to the partition π. If T is π-pure, then $\pi_T = \omega_T$ and, therefore, $\mathcal{H}_\beta(\pi_T) = 0$. Thus, the smaller $\mathcal{H}_\beta(\pi_T)$, the more pure the set T is.

The largest value of $\mathcal{H}_\beta(\pi)$ when $\pi \in \mathsf{PART}(S)$ is achieved when $\pi = \iota_S$; in this case we have:

$$\mathcal{H}_\beta(\iota_S) = \frac{1}{2^{1-\beta} - 1} \left(\frac{1}{|S|^{\beta-1}} - 1\right).$$

Axiom **(P3)** can be extended as follows.

Theorem 2.1 *Let $S_1, \ldots, S_n$ be n pairwise disjoint finite sets, $S = \bigcup_{i=1}^n S_i$ and let $\pi_1, \ldots, \pi_n$ be partitions of $S_1, \ldots, S_n$, respectively.*

We have:

$$\mathcal{H}_\beta(\pi_1 + \cdots + \pi_n) = \sum_{i=1}^n \left(\frac{|S_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i) + \mathcal{H}_\beta(\theta),$$

where θ is the partition $\{S_1, \ldots, S_n\}$ of S.

Proof. The argument is by induction on $n \geq 2$. The basis case, $n = 2$, reduces to Axiom **(P3)**. Suppose that the statement holds for $n - 1$ and let

$T = \bigcup_{i=1}^{n-2}(S)$. We have $\pi_1 + \cdots + \pi_{n-1} \in \mathsf{PART}(T)$ and

$$\begin{aligned}
&\mathcal{H}_\beta(\pi_1 + \cdots + \pi_{n-1} + \pi_n) \\
&= \mathcal{H}_\beta((\pi_1 + \cdots + \pi_{n-1}) + \pi_n) \\
&= \left(\frac{|T|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_1 + \cdots + \pi_{n-1}) + \left(\frac{|S_n|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_n) + \mathcal{H}_\beta(\{T, S_n\}) \\
&\quad \text{(by Axiom } \textbf{(P3)}\text{)} \\
&= \left(\frac{|T|}{|S|}\right)^\beta \left(\sum_{i=1}^{n-1} \left(\frac{|S_i|}{|T|}\right)^\beta \mathcal{H}_\beta(\pi_i) + \mathcal{H}_\beta(\theta')\right) + \left(\frac{|S_n|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_n) + \mathcal{H}_\beta(\{T, S_n\}) \\
&\quad \text{(by the inductive hypothesis),}
\end{aligned}$$

where $\theta' = \{S_1, \ldots, S_{n-1}\} \in \mathsf{PART}(T)$. Note that $\theta = \theta' + \pi_n$. Therefore, we have:

$$\begin{aligned}
&\mathcal{H}_\beta(\pi_1 + \cdots + \pi_n) \\
&= \sum_{i=1}^{n} \left(\frac{|S_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i) + \left(\frac{|T|}{|S|}\right)^\beta \mathcal{H}_\beta(\theta') + \left(\frac{|S_n|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_n) + \mathcal{H}_\beta(\{T, S_n\}) \\
&= \sum_{i=1}^{n} \left(\frac{|S_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i) + \mathcal{H}_\beta(\theta' + \pi_n) \\
&= \sum_{i=1}^{n} \left(\frac{|S_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i) + \mathcal{H}_\beta(\theta).
\end{aligned}$$

∎

3 Conditional β-Entropy of Partitions and Metrics on Partitions

The β-entropy defines naturally a conditional entropy of partitions. We note that the definition introduced here is an improvement over our previous definition given in [SJ02]. Starting from conditional entropies we will be able to define a family of metrics on the set of partitions of a finite set and study the geometry of these finite metric spaces.

Definition 3.1 Let $\pi, \sigma \in \mathsf{PART}(S)$ and let $\sigma = \{C_1, \ldots, C_n\}$. The *$\beta$-conditional entropy* of the partitions $\pi, \sigma \in \mathsf{PART}(S)$ is the function $\mathcal{H} : \mathsf{PART}(S)^2 \longrightarrow \mathbb{R}_{\geq 0}$ defined by:

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

□

Observe that $\mathcal{H}_\beta(\pi|\omega_S) = \mathcal{H}_\beta(\pi)$ and that $\mathcal{H}_\beta(\omega_S|\pi) = \mathcal{H}_\beta(\pi|\iota_S) = 0$ for every partition $\pi \in \mathsf{PART}(S)$. Also, we can write:

$$\mathcal{H}_\beta(\iota_S|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\iota_{C_j}) = \frac{1}{2^{1-\beta}-1}\left(\frac{1}{|S|^{\beta-1}} - \sum_{j=1}^{n}\left(\frac{|C_j|}{|S|}\right)^\beta\right), \quad (1)$$

where $\sigma = \{C_1, \ldots, C_n\}$. The conditional entropy can be written explicitly as:

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

where $\pi = \{B_1, \ldots, B_m\}$.

Theorem 3.2 *Let π, σ be two partitions of a finite set S. 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 $1 \leq j \leq n$ and, therefore,

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

Conversely, suppose that

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

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

The next statement is a generalization of a well-known property of Shannon's entropy.

Theorem 3.3 *Let π, σ be two partitions of a finite set S. We have:*

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

Proof. Suppose that $\pi = \{B_1, \ldots, B_m\}$ and that $\sigma = \{C_1, \ldots, C_n\}$. Observe that

$$\pi \wedge \sigma = \pi_{C_1} + \cdots + \pi_{C_n} = \sigma_{B_1} + \cdots + \sigma_{B_m}.$$

Therefore, by Theorem 2.1 we have:

$$\mathcal{H}_\beta(\pi \wedge \sigma) = \sum_{j=1}^{n}\left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_i}) + \mathcal{H}_\beta(\sigma),$$

which implies

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

The second equality has a similar proof. ∎

The β-conditional entropy is dually monotonic with respect to its first argument and is monotonic with respect to its second argument, as we show in the following statement:

Theorem 3.4 *Let $\pi, \sigma, \sigma' \in \mathsf{PART}(S)$, where S is a finite set. If $\sigma \leq \sigma'$, then $\mathcal{H}_\beta(\sigma|\pi) \geq \mathcal{H}_\beta(\sigma'|\pi)$ and $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi|\sigma')$.*

Proof. Since $\sigma \leq \sigma'$ we have $\pi \wedge \sigma \leq \pi \wedge \sigma'$, so $\mathcal{H}_\beta(\pi \wedge \sigma) \geq \mathcal{H}_\beta(\pi \wedge \sigma')$ by Axiom **(P1)**. Therefore,

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

by Theorem 3.3, which implies $\mathcal{H}_\beta(\sigma|\pi) \geq \mathcal{H}_\beta(\sigma'|\pi)$.

For the second part of the theorem it suffices to prove the inequality for partitions σ, σ' such that $\sigma \prec \sigma'$. Without restricting the generality we may assume that $\sigma = \{C_1, \ldots, C_{n-2}, C_{n-1}, C_n\}$ and $\sigma' = \{C_1, \ldots, C_{n-2}, C_{n-1} \cup C_n\}$. Thus, we can write:

$$\begin{aligned}
&\mathcal{H}_\beta(\pi|\sigma') \\
&= \sum_{i=1}^{n-2} \left(\frac{|C_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_i}) + \left(\frac{|C_{n-1} \cup C_n|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_{n-1} \cup C_n}) \\
&= \sum_{i=1}^{n-2} \left(\frac{|C_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_i}) + \left(\frac{|C_{n-1}| + |C_n|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_{n-1}} + \pi_{C_n}).
\end{aligned}$$

By Axiom **(P3)** we can further write:

$$\begin{aligned}
&\mathcal{H}_\beta(\pi|\sigma') \\
&= \sum_{i=1}^{n-2} \left(\frac{|C_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_i}) + \left(\frac{|C_{n-1}| + |C_n|}{|S|}\right)^\beta \cdot \\
&\quad \left[\left(\frac{|C_{n-1}|}{|C_{n-1}| + |C_n|}\right)^\beta \mathcal{H}_\beta(\pi_{C_{n-1}}) + \left(\frac{|C_n|}{|C_{n-1}| + |C_n|}\right)^\beta \mathcal{H}_\beta(\pi_{C_n}) + \right. \\
&\quad \left. \mathcal{H}_\beta(\{C_{n-1}, C_n\})\right] \\
&= \mathcal{H}_\beta(\pi|\sigma) + \left(\frac{|C_{n-1}| + |C_n|}{|S|}\right)^\beta \mathcal{H}_\beta(\{C_{n-1}, C_n\}) \\
&\geq \mathcal{H}_\beta(\pi|\sigma).
\end{aligned}$$

∎

Corollary 3.5 *Since $\mathcal{H}_\beta(\pi) = \mathcal{H}_\beta(\pi|\omega_S)$ it follows that if $\pi, \sigma \in \mathsf{PART}(S)$, then $\mathcal{H}_\beta(\pi) \geq \mathcal{H}_\beta(\pi|\sigma)$.*

Proof. We observed that $\mathcal{H}_\beta(\pi) = \mathcal{H}_\beta(\pi|\omega_S)$. By the second part of Theorem 3.4, $\omega_S \geq \sigma$ yields the desired inequality. ∎

The next statement that follows from the previous theorem is useful in Section 5.

Corollary 3.6 *Let ξ, θ, θ' be three partitions of a finite set S. If $\theta \geq \theta'$, then*

$$\mathcal{H}_\beta(\xi \wedge \theta) - \mathcal{H}_\beta(\theta) \geq \mathcal{H}_\beta(\xi \wedge \theta') - \mathcal{H}_\beta(\theta').$$

Proof. By Theorem 3.3 we have:

$$\mathcal{H}_\beta(\xi \wedge \theta) - \mathcal{H}_\beta(\xi \wedge \theta') = \mathcal{H}_\beta(\xi|\theta) + \mathcal{H}_\beta(\theta) - \mathcal{H}_\beta(\xi|\theta') - \mathcal{H}_\beta(\theta').$$

The monotonicity of $\mathcal{H}_\beta(|)$ in its second argument means that: $\mathcal{H}_\beta(\xi|\theta) - \mathcal{H}_\beta(\xi|\theta') \geq 0$, so $\mathcal{H}_\beta(\xi \wedge \theta) - \mathcal{H}_\beta(\xi \wedge \theta') \geq \mathcal{H}_\beta(\theta) - \mathcal{H}_\beta(\theta')$, which implies the desired inequality. ∎

The behavior of β-conditional entropies with respect to the "addition" of partitions is discussed in the next statement.

Theorem 3.7 *Let S be a finite set, π, θ be two partitions of S, where $\theta = \{D_1, \ldots, D_h\}$. If $\sigma_i \in \mathsf{PART}(D_i)$ for $1 \leq i \leq h$, then*

$$\mathcal{H}_\beta(\pi|\sigma_1 + \cdots + \sigma_h) = \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{D_i}|\sigma_i).$$

If $\tau = \{F_1, \ldots, F_k\}$, $\sigma = \{C_1, \ldots, C_n\}$ be two partitions of S, and let $\pi_i \in \mathsf{PART}(F_i)$ for $1 \leq i \leq k$. Then,

$$\mathcal{H}_\beta(\pi_1 + \cdots + \pi_k|\sigma) = \sum_{i=1}^{k} \left(\frac{|F_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i|\sigma_{F_i}) + \mathcal{H}_\beta(\tau|\sigma).$$

Proof. Suppose that $\sigma_i = \{E_i^\ell \mid 1 \leq \ell \leq p_i\}$. The blocks of the partition $\sigma_1 + \cdots + \sigma_h$ are the sets of the collection $\bigcup_{i=1}^{h}\{E_i^\ell \mid 1 \leq \ell \leq p_i\}$. Thus, we have:

$$\mathcal{H}_\beta(\pi|\sigma_1 + \cdots + \sigma_h) = \sum_{i=1}^{h}\sum_{\ell=1}^{p_i} \left(\frac{|E_i^\ell|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{E_i^\ell}).$$

On the other hand, since $(\pi_{D_i})_{E_i^\ell} = \pi_{E_i^\ell}$, we have:

$$\begin{aligned}\sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{D_i}|\sigma_i) &= \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta \sum_{\ell=1}^{p_i} \left(\frac{|E_i^\ell|}{|D_i|}\right)^\beta \mathcal{H}_\beta(\pi_{E_i^\ell}) \\ &= \sum_{i=1}^{h}\sum_{\ell=1}^{p_i} \left(\frac{|E_i^\ell|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{E_i^\ell}),\end{aligned}$$

which gives the first equality of the theorem.

To prove the second part observe that $(\pi_1+\cdots+\pi_k)_{C_j} = (\pi_1)_{C_j}+\cdots+(\pi_k)_{C_j}$ for every block C_j of σ. Thus, we have:

$$\mathcal{H}_\beta((\pi_1 + \cdots + \pi_k|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta((\pi_1)_{C_j} + \cdots + (\pi_k)_{C_j}).$$

By applying Theorem 2.1 to partitions $(\pi_1)_{C_j}, \ldots, (\pi_k)_{C_j}$ of C_j we can write:

$$\mathcal{H}_\beta((\pi_1)_{C_j} + \cdots + (\pi_k)_{C_j}) = \sum_{i=1}^{k} \left(\frac{|F_i \cap C_j|}{|C_j|}\right)^\beta \mathcal{H}_\beta((\pi_i)_{C_j}) + \mathcal{H}_\beta(C_j).$$

Thus,

$$\begin{aligned}
&\mathcal{H}_\beta((\pi_1)_{C_j} + \cdots + (\pi_k)_{C_j}) \\
&= \sum_{j=1}^{n}\sum_{i=1}^{k} \left(\frac{|F_i \cap C_j|}{|S|}\right)^\beta \mathcal{H}_\beta((\pi_i)_{C_j}) + \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\tau_{C_j}) \\
&= \sum_{i=1}^{k} \left(\frac{|F_i|}{|S|}\right)^\beta \sum_{j=1}^{n} \left(\frac{|F_i \cap C_j|}{|F_i|}\right)^\beta \mathcal{H}_\beta((\pi_i)_{F_i \cap C_j}) + \mathcal{H}_\beta(\tau|\sigma) \\
&= \sum_{i=1}^{k} \left(\frac{|F_i|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_i|\sigma_{F_i}) + \mathcal{H}_\beta(\tau|\sigma),
\end{aligned}$$

which is the desired equality. ∎

In [dM91] L. de Mántaras proved that Shannon's entropy generates a metric $d : \mathsf{PART}(S)^2 \longrightarrow \mathbb{R}^2$ given by $d(\pi, \sigma) = \mathcal{H}(\pi|\sigma) + \mathcal{H}(\sigma|\pi)$, for $\pi, \sigma \in \mathsf{PART}(S)$. We extend his result to a class of metrics that can be defined by β-entropies, thereby improving our earlier results [SJ03]. To this end we need the following statement:

Theorem 3.8 *Let π, σ, τ be three partitions of the finite set S. We have:*

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

Proof. Suppose that $\sigma = \{C_1, \ldots, C_n\}$ and $\tau = \{D_1, \ldots, D_p\}$. We observed already that

$$\sigma \wedge \tau = \sigma_{D_1} + \cdots + \sigma_{D_p} = \tau_{C_1} + \cdots + \tau_{C_n}.$$

Consequently, by Theorem 3.7, we have

$$\begin{aligned}
\mathcal{H}_\beta(\pi|\sigma \wedge \tau) &= \mathcal{H}_\beta(\pi|\sigma_{D_1} + \cdots + \sigma_{D_p}) \\
&= \sum_{l=1}^{p} \left(\frac{|D_l|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{D_l}|\sigma_{D_l}).
\end{aligned}$$

Also, we have

$$\mathcal{H}_\beta(\sigma|\tau) = \sum_{l=1}^{p} \left(\frac{|D_l|}{|S|}\right)^\beta \mathcal{H}_\beta(\sigma_{D_l}).$$

The last two equalities imply:

$$\begin{aligned}\mathcal{H}_\beta(\pi|\sigma\wedge\tau)+\mathcal{H}_\beta(\sigma|\tau) &= \sum_{l=1}^{p}\left(\frac{|D_l|}{|S|}\right)^\beta(\mathcal{H}_\beta(\pi_{D_l}|\sigma_{D_l})+\mathcal{H}_\beta(\sigma_{D_l}))\\ &= \sum_{l=1}^{p}\left(\frac{|D_l|}{|S|}\right)^\beta\mathcal{H}_\beta(\pi_{D_l}\wedge\sigma_{D_l})\\ &\quad\text{(by Theorem 3.3)}\\ &= \sum_{l=1}^{p}\left(\frac{|D_l|}{|S|}\right)^\beta\mathcal{H}_\beta((\pi\wedge\sigma)_{D_l})\\ &= \mathcal{H}_\beta(\pi\wedge\sigma|\tau),\end{aligned}$$

which is the equality we seek to prove. ∎

Corollary 3.9 *Let π,σ,τ be three partitions of the finite set S. Then, we have:*

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

Proof. By Theorem 3.8, the monotonicity of β-conditional entropy in its second argument and the dual monotonicity of the same in its first argument we can write:

$$\begin{aligned}\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\tau) &\geq \mathcal{H}_\beta(\pi|\sigma\wedge\tau)+\mathcal{H}_\beta(\sigma|\tau)\\ &= \mathcal{H}_\beta(\pi\wedge\sigma|\tau)\\ &\geq \mathcal{H}_\beta(\pi|\tau),\end{aligned}$$

which is the desired inequality. ∎

We can show now a central result:

Theorem 3.10 *The mapping $d_\beta:\mathsf{PART}(S)^2\longrightarrow\mathbb{R}_{\geq 0}$ defined by: $d_\beta(\pi,\sigma)=\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\pi)$ for $\pi,\sigma\in\mathsf{PART}(S)$ is a metric on $\mathsf{PART}(S)$.*

Proof. A double application of Corollary 3.9 yields:

$$\begin{aligned}\mathcal{H}_\beta(\pi|\sigma)+\mathcal{H}_\beta(\sigma|\tau) &\geq \mathcal{H}_\beta(\pi|\tau),\\ \mathcal{H}_\beta(\sigma|\pi)+\mathcal{H}_\beta(\tau|\sigma) &\geq \mathcal{H}_\beta(\tau|\pi).\end{aligned}$$

Adding these inequality gives

$$d_\beta(\pi,\sigma)+d_\beta(\sigma,\tau)\geq d_\beta(\pi,\tau),$$

which is the triangular inequality for d_β.

The symmetry of d_β is obvious and it is clear that $d_\beta(\pi,\pi)=0$ for every $\pi\in\mathsf{PART}(S)$.

Suppose now that $d_\beta(\pi,\sigma)=0$. Since the values of β-conditional entropies are non-negative this implies $\mathcal{H}_\beta(\pi|\sigma)=\mathcal{H}_\beta(\sigma|\pi)=0$. By Theorem 3.2 we

have both $\sigma \leq \pi$ and $\pi \leq \sigma$, respectively, so $\pi = \sigma$. Thus, d_β is a metric on $\mathsf{PART}(S)$. ∎

It is clear that $d_\beta(\pi, \omega_S) = \mathcal{H}_\beta(\pi)$ and $d_\beta(\pi, \iota_S) = \mathcal{H}(\iota_S|\pi)$.

The behavior of the distance d_β with respect to partition addition is discussed in the next statement.

Theorem 3.11 *Let S be a finite set, π, θ be two partitions of S, where $\theta = \{D_1, \ldots, D_h\}$. If $\sigma_i \in \mathsf{PART}(D_i)$ for $1 \leq i \leq h$, then*

$$d_\beta(\pi, \sigma_1 + \cdots + \sigma_h) = \sum_{i=1}^{h} \left(\frac{|D_i|}{|S|}\right)^\beta d_\beta(\pi_{D_i}, \sigma_i) + \mathcal{H}_\beta(\theta|\pi).$$

Proof. The theorem follows directly from Theorems 3.7 and **??**. ∎

4 The Metric Geometry of the Partition Space

The distance between two partitions can be expressed using distances relative to the total partition or to the identity partition. Indeed, note that for $\pi, \sigma \in \mathsf{PART}(S)$, where $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$ we have:

$$d_\beta(\pi, \sigma) = \frac{1}{(2^{1-\beta} - 1)|S|^\beta} \left(2 \sum_{i=1}^{n} \sum_{j=1}^{m} |B_i \cap C_j|^\beta - \sum_{i=1}^{n} |B_i|^\beta - \sum_{j=1}^{m} |C_j|^\beta\right), \quad (3)$$

In the special case, when $\sigma = \omega$ we have:

$$\begin{aligned} d_\beta(\pi, \omega) &= \frac{1}{(2^{1-\beta} - 1)|S|^\beta} \left(2 \sum_{i=1}^{n} |B_i|^\beta - \sum_{i=1}^{n} |B_i|^\beta - |S|^\beta\right) \\ &= \frac{1}{(2^{1-\beta} - 1)|S|^\beta} \left(\sum_{i=1}^{n} |B_i|^\beta - |S|^\beta\right). \end{aligned}$$

Similarly, we can write:

$$d(\iota, \sigma) = \frac{1}{(2^{1-\beta} - 1)|S|^\beta} \left(|S| - \sum_{j=1}^{n} |C_j|^\beta\right).$$

We have the following result:

Theorem 4.1 *Let $\pi, \sigma \in \mathsf{PART}(S)$ be two partitions. We have:*

$$\begin{aligned} d_\beta(\pi, \sigma) &= 2 \cdot d_\beta(\pi \wedge \sigma, \omega_S) - d_\beta(\pi, \omega_S) - d_\beta(\sigma, \omega_S) \\ &= d_\beta(\iota_S, \pi) + d_\beta(\iota_S, \sigma) - 2 \cdot d_\beta(\iota_S, \pi \wedge \sigma). \end{aligned}$$

Proof. Starting from the expression of the distance we can write:

$$\begin{aligned}
&d_\beta(\pi,\sigma)+d_\beta(\pi,\omega_S)+d_\beta(\sigma,\omega_S)\\
&= \frac{1}{(2^{1-\beta}-1)|S|^\beta}\left(2\sum_{i=1}^{n}\sum_{j=1}^{m}|B_i\cap C_j|^\beta-\sum_{i=1}^{n}|B_i|^\beta-\sum_{j=1}^{m}|C_j|^\beta\right)+\\
&\quad \frac{1}{(2^{1-\beta}-1)|S|^\beta}\left(\sum_{i=1}^{n}|B_i|^\beta-|S|^\beta\right)+\frac{1}{(2^{1-\beta}-1)|S|^\beta}\left(\sum_{j=1}^{m}|C_j|^\beta-|S|^\beta\right)\\
&= \frac{2}{(2^{1-\beta}-1)|S|^\beta}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}|B_i\cap C_j|^\beta-|S|^\beta\right)\\
&= 2\cdot d_\beta(\pi\wedge\sigma,\omega_S).
\end{aligned}$$

The proof of the second equality is similar and is omitted. ∎

Corollary 4.2 *Let θ,τ be two partitions from $\mathsf{PART}(S)$. If $\theta\leq\tau$ and we have either $d_\beta(\theta,\omega_S)=d_\beta(\tau,\omega_S)$ or $d_\beta(\iota_S,\theta)=d_\beta(\iota_S,\tau)$, then $\theta=\tau$.*

Proof. Observe that if $\theta\leq\tau$, then Theorem 4.1 implies

$$d_\beta(\theta,\tau)+d_\beta(\tau,\omega_S)=d_\beta(\theta,\omega_S),$$

and

$$d_\beta(\theta,\tau)=d_\beta(\iota_S,\tau)-d_\beta(\iota_S,\theta).$$

Suppose that $d_\beta(\theta,\omega_S)=d_\beta(\tau,\omega_S)$. Since $d_\beta(\tau,\omega_S)=d_\beta(\theta,\omega_S)$ it follows that $d_\beta(\theta,\tau)=0$, so $\theta=\tau$.

If $d_\beta(\iota_S,\theta)=d_\beta(\iota_S,\tau)$ the same conclusion can be reached immediately. ∎

Theorem 4.3 *Let $\pi,\sigma\in\mathsf{PART}(S)$. The following statements are equivalent:*

1. *$\sigma\leq\pi$;*
2. *we have $[\sigma,\pi,\omega_S]$ in the metric space $(\mathsf{PART}(S),d_\beta)$;*
3. *we have $[\iota_S,\sigma,\pi]$ in the metric space $(\mathsf{PART}(S),d_\beta)$.*

Proof. We prove that Part (1) implies both Parts (2) and (3). Suppose that $\sigma\leq\pi$. By Theorem 4.1, since $\sigma\wedge\pi=\sigma$ we have both

$$\begin{aligned}
d_\beta(\pi,\sigma) &= d_\beta(\sigma,\omega_S)-d_\beta(\pi,\omega_S)\\
&= d_\beta(\iota_S,\pi)-d_\beta(\iota_S,\sigma),
\end{aligned}$$

which are equivalent to $[\sigma,\pi,\omega_S]$ and $[\iota_S,\sigma,\pi]$, respectively.

Conversely, suppose that $[\pi,\sigma,\omega_S]$, that is,

$$d_\beta(\pi,\sigma)+d_\beta(\sigma,\omega_S)=d_\beta(\pi,\omega_S).$$

Theorem 4.1 implies $d_\beta(\pi, \omega_S) = d_\beta(\pi \wedge \sigma, \omega_S)$. Since $\pi \wedge \sigma \leq \pi$, by Corollary 4.2, we have $\pi = \pi \wedge \sigma$, so $\pi \leq \sigma$. Thus, the second statement of the theorem implies the first. Finally, the betweenness $[\iota_S, \pi, \sigma]$ means that $d_\beta(\iota_S, \pi) = d_\beta(\iota_S, \sigma) + d_\beta(\sigma, \pi)$, which implies $d_\beta(\iota_S, \sigma) = d_\beta(\iota_S, \pi \wedge \sigma)$. By the same Corollary 4.2 we obtain the equality $\sigma = \pi \wedge \sigma$, so $\sigma \leq \pi$. This shows that the third statement implies the first. ∎

Metrics generated by β-conditional entropies are closely related to lower valuations of the upper semi-modular lattices of partitions of finite sets. This connection was established in [Bir73] and studied in [BL95, Bar78, Mon81].

A *lower valuation* on a lattice $(L, \vee, \wedge)$ is a mapping $v : L \longrightarrow \mathbb{R}$ such that $v(\pi \vee \sigma) + v(\pi \wedge \sigma) \geq v(\pi) + v(\sigma)$ for every $\pi, \sigma \in L$. If the reverse inequality is satisfied, that is, if $v(\pi \vee \sigma) + v(\pi \wedge \sigma) \leq v(\pi) + v(\sigma)$ for every $\pi, \sigma \in L$, then v is referred to as an *upper valuation*.

If $v \in L$ is both a lower and upper valuation, that is, if $v(\pi \vee \sigma) + v(\pi \wedge \sigma) = v(\pi) + v(\sigma)$ for every $\pi, \sigma \in L$, then v is a valuation on L. It is known [Bir73] that if there exists a positive valuation v on L, then L must be a modular lattice. Since the partition lattice of a set is an upper-semimodular lattice that is not modular ([Bir73]) it is clear that positive valuations do not exist on partition lattices. However, lower and upper valuations do exist, as shown next:

Theorem 4.4 *Let S be a finite set. Define the mappings $v_\beta : \mathsf{PART}(S) \longrightarrow \mathbb{R}$ and let $w_\beta : \mathsf{PART}(S) \longrightarrow \mathbb{R}$ be by $v_\beta(\pi) = d_\beta(\iota_S, \pi)$ and $w_\beta(\pi) = d_\beta(\pi, \omega_S)$, respectively, for $\pi \in \mathsf{PART}(S)$. Then, v_β is a lower valuation and w_β is an upper valuation on the lattice $(\mathsf{PART}(S), \vee, \wedge)$.*

Proof. Theorem 4.1 allows us to write:

$$\begin{aligned} d_\beta(\pi, \sigma) &= v_\beta(\pi) + v_\beta(\sigma) - 2v_\beta(\pi \wedge \sigma) \\ &= 2w_\beta(\pi \wedge \sigma) - w_\beta(\pi) - w_\beta(\sigma), \end{aligned}$$

for every $\pi, \sigma \in \mathsf{PART}(S)$.

If we rewrite the triangular inequality $d_\beta(\pi, \tau) + d_\beta(\tau, \sigma) \geq d_\beta(\pi, \sigma)$ using the valuations v_β and w_β we obtain:

$$\begin{aligned} v_\beta(\tau) + v_\beta(\pi \wedge \sigma) &\geq v_\beta(\pi \wedge \tau) + v_\beta(\tau \wedge \sigma), \\ w_\beta(\pi \wedge \tau) + w_\beta(\tau \wedge \sigma) &\geq w_\beta(\tau) + w_\beta(\pi \wedge \sigma), \end{aligned}$$

for every $\pi, \tau, \sigma \in \mathsf{PART}(S)$. If we choose $\tau = \pi \vee \sigma$ the last inequalities yield:

$$\begin{aligned} v_\beta(\pi) + v_\beta(\sigma) &\leq v_\beta(\pi \vee \sigma) + v_\beta(\pi \wedge \sigma) \\ w_\beta(\pi) + w_\beta(\sigma) &\geq w_\beta(\pi \vee \sigma) + w_\beta(\pi \wedge \sigma), \end{aligned}$$

for every $\pi, \sigma \in \mathsf{PART}(S)$, which shows that v_β is a lower valuation and w_β is an upper valuation on the lattice $(\mathsf{PART}(S), \vee, \wedge)$. ∎

5 Metrics and Data Mining

We begin by defining the notion of *object system* as a triple $\mathcal{S} = (S, H, C)$, where S is a finite set referred to as the *training set*, $H = \{A_1, \ldots, A_n\}$ is a finite set of mappings of the form $A_i : S \longrightarrow D_i$ called the *features of* $\mathcal{S}$ for $1 \leq i \leq n$, and $C : S \longrightarrow D$ is the *classification function*. The sets $D_1, \ldots, D_n$ are supposed to contain at least two elements and they are referred as the *domains of the attributes* $A_1, \ldots, A_n$.

A set of attributes X, $X \subseteq H$ generates a mapping $\wp_X : S \longrightarrow \bigcup'\{D_i \mid A_i \in X\}$, defined by $\wp_X(t) = \{(A(t), A) \mid A \in X\}$ for every $t \in S$, where $\bigcup'$ denotes the disjoint union of a family of sets; we refer to $\wp_X$ as the *projection on* X of $\mathcal{S}$. Projections define partitions on the set of objects in a natural manner; namely if X is a set of attributes, a block B_v of the partition π^X is a non-empty set of the form $\{t \in S | \wp_X(t) = v\}$, where v is an element of the range of $\wp_X$.

To introduce formally the notion of decision tree we start from the notion of tree domain. A *tree domain* is a non-empty set of sequences D over the set of natural numbers $\mathbb{N}$ that satisfies the following conditions:

1. every prefix of a sequence $s \in D$ also belongs to D, and
2. for every $m \geq 1$, if $(p_1, \ldots, p_{m-1}, p_m) \in D$, then $(p_1, \ldots, p_{m-1}, q) \in D$ for every $q \leq p_m$.

The elements of D are called the *vertices* of D. If u and v are vertices of D and u is a prefix of v, then we refer to v as a *descendant* of u and to u as an *ancestor* of v. If $v = ui$ for some $i \in \mathbb{N}$, then we call v an *immediate descendant* of u and u an *immediate ancestor* of v. The *root* of every tree domain is the null sequence λ. A *leaf* of D is a vertex of D with no immediate descendants.

Let S be a finite set and let D be a tree domain. An S-tree is a function $\mathcal{T} : D \longrightarrow \mathcal{P}(S)$ such that $\mathcal{T}(\lambda) = S$, and if $u1, \ldots, um$ are the descendants of a vertex u, then the sets $\mathcal{T}(u1), \ldots, \mathcal{T}(um)$ form a partition of the set $\mathcal{T}(u)$.

A *decision tree* for an object system $\mathcal{S} = (S, H, C)$ is an S-tree $\mathcal{T}$, such that if the vertex v has the descendants $v0, \ldots, vm$, then there exists an attribute $A \in H$ (called the *splitting attribute* in v) such that $\{\mathcal{T}(vi) \mid 1 \leq i \leq m\}$ is the partition $\pi^A_{\mathcal{T}(v)}$.

Thus, each descendant vi of a vertex v corresponds to a value a of the attribute A that was used as a splitting attribute in v. If $\lambda = v_1, v_2, \ldots, v_k = u$ is the path in $\mathcal{T}$ that was used to reach the vertex u, $A_{i_1}, A_{i_2}, \ldots, A_{i_{k-1}}$ are the splitting attributes in $v_0, v_1, \ldots, v_{k-1}$ and $a_1, a_2, \ldots, a_{k-1}$ are the values that correspond to $v_2, \ldots, v_k$, respectively, then we say that u is reached by the selection:

$$A_{i_1} = a_1 \wedge \cdots \wedge A_{i_{k-1}} = a_{k-1}.$$

It is desirable that the leaves of a decision tree contain C-pure or almost C-pure sets of objects. In other words, the objects assigned to a leaf of the tree should, with few exceptions, have the the same value for the class attribute C. This amounts to asking that for each leaf w of $\mathcal{T}$ we must have $\mathcal{H}_\beta(\pi^C_{S_w})$ as

close to 0 as possible. To take into account the size of the leaves note that the collection of sets of objects assigned to the leafs is a partition κ of S and that we need to minimize:

$$\sum_w \left(\frac{|S_w|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi^C_{S_w}),$$

which is the conditional entropy $\mathcal{H}(\pi^C|\kappa)$. By Theorem 3.2 we have $\mathcal{H}(\pi^C|\kappa) = 0$ if and only if $\kappa \leq \pi^C$, which happens when the sets of objects assigned to the leafs are C-pure.

The construction of a decision tree $\mathcal{T}_\beta(\mathcal{S})$ for an object system $\mathcal{S} = (S, H, C)$ evolves in a top-down manner according to the following high-level description of a general algorithm [TSK05]. The algorithm starts with an object system $\mathcal{S} = (S, H, C)$, a value of β and with an impurity threshold ϵ and it consists of the following steps:

1. If $\mathcal{H}_\beta(\pi^C_S) \leq \epsilon$, then return $\mathcal{T}$ as an one-vertex tree; otherwise go to 2.

2. Assign the set S to a vertex v, choose an attribute A as a splitting attribute of S (using a splitting attribute criterion to be discussed in the sequel) and apply the algorithm to the object systems $(S_{a_1}, H, C), \ldots, (S_{a_p}, H, C)$, where $S_{a_i} = \{t \in S \mid A(t) = a_i\} \neq \emptyset$. Let $\mathcal{T}_1, \ldots, \mathcal{T}_p$ the decision trees returned for the systems $\mathcal{S}_1, \ldots, \mathcal{S}_p$, respectively. Connect the roots of these trees to v.

Note that if ϵ is sufficiently small and if $\mathcal{H}_\beta(\pi^C_S) \leq \epsilon$, where $S = \mathcal{T}(u)$ is the set of objects at a node u, then there is a block Q_k of the partition π^C_S that is dominant in the set S. We refer to Q_k as the dominant class of u.

Once a decision tree $\mathcal{T}$ is built it can be used to determine the class of a new object $t \notin S$ such that the attributes of the set H are applicable. If $A_{i_1}(t) = a_1, \ldots, A_{i_{k-1}}(t) = a_{k-1}$, a leaf u was reached through the path $v_1, \ldots, v_k = u$, and $a_1, a_2, \ldots, a_{k-1}$ are the values that correspond to $v_2, \ldots, v_k$, respectively, then t is classified in the class Q_k, where Q_k is the dominant class at leaf u.

The description of the algorithm shows that the construction of a decision tree depends essentially on the method for choosing the splitting attribute. We focus next on this issue.

Classical decision tree algorithms make use of the information gain criterion or the gain ratio to choose splitting attribute. These criteria are formulated using Shannon's entropy, as their designations indicate.

In our terms, the analogue of the information gain for a vertex w and an attribute A is: $\mathcal{H}_\beta(\pi^C_{S_w}) - \mathcal{H}_\beta(\pi^C_{S_w}|\pi^A_{S_w})$. The selected attribute is the one that realizes the highest value of this quantity. When $\beta \to 1$ we obtain the information gain linked to Shannon entropy. When $\beta = 2$ one obtains the selection criteria for the Gini index using the CART algorithm [BFOS98].

The monotonicity property of conditional entropy shows that if A, B are two attributes such that $\pi^A \leq \pi^B$ (which indicates that the domain of A has more values than the domain of B), then $\mathcal{H}_\beta(\pi^C_{S_w}|\pi^A_{S_w}) \leq \mathcal{H}_\beta(\pi^C_{S_w}|\pi^B_{S_w})$, so the gain for A is larger than the gain for B. This highlights a well-known problem of

choosing attributes based on information gain and related criteria: these criteria favor attributes with large domains, which in turn, generate bushy trees. To alleviate this problem information gain was replaced with the information gain ratio defined as:

$$\frac{\mathcal{H}_\beta(\pi^C_{S_w}) - \mathcal{H}_\beta(\pi^C_{S_w}|\pi^A_{S_w})}{\mathcal{H}_\beta(\pi^A_{S_w})},$$

which introduces the compensating divisor $\mathcal{H}_\beta(\pi^A_{S_w})$.

We propose replacing the information gain and the gain ratio criteria by choosing as splitting attribute for a node w an attribute that minimizes the distance $d_\beta(\pi^C_{S_w}, \pi^A_{S_w}) = \mathcal{H}_\beta(\pi^C_{S_w}|\pi^A_{S_w}) + \mathcal{H}_\beta(\pi^A_{S_w}|\pi^C_{S_w})$. This idea has been developed by L. de Mántaras in [dM91] for the metric d_1 induced by Shannon's entropy. Since one could obtain better classifiers for various data sets and user needs using values of β that are different from one, our approach is an improvement of previous results.

Besides being geometrically intuitive, the minimal distance criterion has the advantage of limiting both conditional entropies $\mathcal{H}_\beta(\pi^C_{S_w}|\pi^A_{S_w})$ and $\mathcal{H}_\beta(\pi^A_{S_w}|\pi^C_{S_w})$. The first limitation insures that the choice of the splitting attribute will provide a high information gain; the second limitation insures that attributes with large domains are not favored over attributes with smaller domains.

Suppose that in the process of building a decision tree for an object system $\mathcal{S} = (S, H, C)$ we constructed a stump of the tree $\mathcal{T}$ that has m leaves and that the sets of objects that correspond to these leaves are $S_1, \ldots, S_n$. This means that we created the partition $\kappa = \{S_1, \ldots, S_n\} \in \mathsf{PART}(S)$, so $\kappa = \omega_{S_1} + \cdots + \omega_{S_n}$. We choose to split the node v_i using as splitting attribute the attribute A that minimizes the distance $d_\beta(\pi^C_{S_i}, \pi^A_{S_i})$. The new partition κ' that replaces κ is

$$\kappa' = \omega_{S_1} + \cdots + \omega_{S_{i-1}} + \pi^A_{S_i} + \omega_{S_{i+1}} + \cdots + \omega_{S_n}.$$

Note that $\kappa \geq \kappa'$. Therefore, we have:

$$\begin{aligned}
d_\beta(\pi^C \wedge \kappa, \kappa) &= d_\beta(\pi^C \wedge \kappa, \omega_S) - d_\beta(\kappa, \omega_S) \\
&\quad \text{(because } [\pi^C \wedge \kappa, \kappa, \omega_S]) \\
&= \mathcal{H}_\beta(\pi^C \wedge \kappa) - \mathcal{H}_\beta(\kappa) \\
&\geq \mathcal{H}_\beta(\pi^C \wedge \kappa') - \mathcal{H}_\beta(\kappa') \\
&\quad \text{(by Corollary 3.6)} \\
&= d_\beta(\pi^C \wedge \kappa', \kappa').
\end{aligned}$$

This shows that as the construction of the tree advances the current partition κ gets closer to the partition $\pi^C \wedge \kappa$. More significantly, as the stump of the tree grows, κ gets closer to the class partition π^C. Indeed, by Theorem 3.11 we can write:

$$\begin{aligned}
d_\beta(\pi^C, \kappa) &= d_\beta(\pi^C, \omega_{S_1} + \cdots + \omega_{S_n}) \\
&= \sum_{j=1}^{n} \left(\frac{|S_j|}{|S|}\right)^\beta d_\beta(\pi^C_{S_j}, \omega_{S_j}) + \mathcal{H}_\beta(\theta|\pi^C),
\end{aligned}$$

where $\theta = \{S_1, \ldots, S_n\}$. Similarly, we can write:

$$\begin{aligned} d_\beta(\pi^C, \kappa') &= d_\beta(\pi^C, \omega_{S_1} + \cdots + \omega_{S_{i-1}} + \pi^A_{S_i} + \omega_{S_{i+1}} + \cdots + \omega_{S_n}) \\ &= \sum_{j=1, j\neq i}^{n} \left(\frac{|S_j|}{|S|}\right)^\beta d_\beta(\pi^C_{S_j}, \omega_{S_j}) + \left(\frac{|S_i|}{|S|}\right)^\beta d_\beta(\pi^C_{S_i}, \pi^A_{S_i}) + \mathcal{H}_\beta(\theta|\pi^C). \end{aligned}$$

These equalities imply:

$$\begin{aligned} d_\beta(\pi^C, \kappa) - d_\beta(\pi^C, \kappa') &= \left(\frac{|S_i|}{|S|}\right)^\beta \left(d_\beta(\pi^C_{S_i}, \omega_{S_i}) - d_\beta(\pi^C_{S_i}, \pi^A_{S_i})\right) \\ &= \left(\frac{|S_i|}{|S|}\right)^\beta \left(\mathcal{H}_\beta(\pi^C_{S_i}) - d_\beta(\pi^C_{S_i}, \pi^A_{S_i})\right). \end{aligned}$$

If the choices of the node and the splitting attribute are made such that:

$$\mathcal{H}_\beta(\pi^C_{S_i}) > d_\beta(\pi^C_{S_i}, \pi^A_{S_i}),$$

then the distance between π^C and the current partition κ of the tree stump will decrease. Since the distance between $\pi^C \wedge \kappa$ and κ decreases in any case when the tree is expanded it follows that the "triangle" determined by π^C, $\pi^C \wedge \kappa$, and κ will shrink during the construction of the decision tree.

6 Experimental Results

We tested our approach on a number of data sets from [BM98]. Due to space limitations we included only the results shown in Figure 1 which are fairly typical. Decision trees were constructed using metrics d_β, where β varied between 0.25 and 2.50. Note that for $\beta = 1$ the metric algorithm coincides with the approach of de Mántaras.

In all cases, accurracy was assessed through 10-fold cross-validation. We also built standard decision trees using the J48 technique of the well-known WEKA package [WF05], which yielded the following results:

Standard J4.8

Data Set	accuracy	size	leaves
Audiology	77.88	54	32
Hepatitis	83.87	21	11
Primary-tumor	39.82	88	47
Vote	94.94	7	4

The experimental evidence shows that β can be adapted such that accuracy is comparable, or better than the standard algorithm. The size of the trees and the number of leaves show that the proposed approach to decision trees results consistently in smaller trees with fewer leaves.

Audiology

β	accuracy	size	leaves
2.50	53.54	53	36
2.25	54.42	53	36
2.00	54.87	54	37
1.75	53.10	47	32
1.50	76.99	29	19
1.25	78.32	29	19
1.00	76.99	29	19
0.75	76.99	29	19
0.50	76.99	29	19
0.25	78.76	33	21

Hepatitis

β	accuracy	size	leaves
2.50	81.94	15	8
2.25	81.94	9	5
2.00	81.94	9	5
1.75	83.23	9	5
1.50	84.52	9	5
1.25	84.52	11	6
1.00	85.16	11	6
0.75	85.81	9	5
0.50	83.23	5	3
0.25	82.58	5	3

Primary-tumor

β	accuracy	size	leaves
2.50	34.81	50	28
2.25	35.99	31	17
2.00	37.76	33	18
1.75	36.28	29	16
1.50	41.89	40	22
1.25	42.18	38	21
1.00	42.48	81	45
0.75	41.30	48	27
0.50	43.36	62	35
0.25	44.25	56	32

Vote

β	accuracy	size	leaves
2.50	94.94	7	4
2.25	94.94	7	4
2.00	94.94	7	4
1.75	94.94	7	4
1.50	95.17	7	4
1.25	95.17	7	4
1.00	95.17	7	4
0.75	94.94	7	4
0.50	95.17	9	5
0.25	95.17	9	5

Figure 1: Experimental Results

7 Conclusion and Future Work

We introduced a family of metrics on the set of partitions of a finite set that can be used for a new splitting criterion for building decision trees. In addition to being more intuitive than the classic approach, this criterion results in decision trees that have smaller sizes and fewer leaves than the trees built with standard methods, and have comparable or better accuracy.

The value of β that results in the smallest trees seems to depend on the relative distribution of the class attribute and the values of the feature attributes of the objects. We believe that further investigations should develop numerical characteristics of data sets that allow predicting "optimal" values for β, that is, values that result in the smallest decision trees for data sets.

Another future direction is related to clustering algorithms. Since clusterings of objects can be regarded as partitions, metrics developed for partitions present an interest for the study of the dynamics of clusters, as clusters are formed during incremental algorithms [SSK04], or as data sets evolve.

References

[Bar78] J.P. Barthélemy. Remarques sur les propriétés metriques des ensembles ordonnés. *Math. Sci. hum.*, 61:39–60, 1978.

[BFOS98] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Chapman and Hall, Boca Raton, 1998.

[Bir73] G. Birkhoff. *Lattice Theory*. American Mathematical Society, Providence, 1973.

[BL95] J.P. Barthélemy and B. Leclerc. The median procedure for partitions. In *Partitioning Data Sets*, pages 3–34, Providence, 1995. American Mathematical Society.

[BM98] C. L. Blake and C. J. Merz. *UCI Repository of machine learning databases*. University of California, Irvine, Dept. of Information and Computer Sciences, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

[Dar70] Z. Daróczy. Generalized information functions. *Information and Control*, 16:36–51, 1970.

[dM91] R. López de Mántaras. A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6:81–92, 1991.

[Ler81] I. C. Lerman. *Classification et analyse ordinale des données*. Dunod, Paris, 1981.

[Mon81] B. Monjardet. Metrics on partially ordered sets – a survey. *Discrete Mathematics*, 35:173–184, 1981.

[SJ02] D. A. Simovici and S. Jaroszewicz. An axiomatization of partition entropy. *IEEE Transactions on Information Theory*, 48:2138–2142, 2002.

[SJ03] D. A. Simovici and S. Jaroszewicz. Generalized entropy and decision trees. In *EGC 2003 - Journees francophones d'Extraction et de Gestion de Connaissances*, pages 369–380, Lyon, France, 2003.

[SSK04] D. A. Simovici, N. Singla, and M. Kuperberg. Metric incremental clustering of nominal data. In *Proceedings of ICDM 2004*, pages 523–527, Brighton, UK, 2004.

[TSK05] P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Addison-Wesley, Boston, 2005.

[WF05] I. H. Witten and E. Frank. *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, second edition, 2005.