# A Greedy Algorithm for Supervised Discretization

Richard Butterworth [a] Dan A. Simovici [a,*] Gustavo S. Santos [b]
Lucila Ohno-Machado [b]

[a] *University of Massachusetts at Boston, Department of Computer Science, Boston, Massachusetts 02125, USA*

[b] *Decision Systems Group, Division of Health Sciences and Technology, Harvard and MIT, Brigham and Womens' Hospital, 75 Francis Street, Boston, MA 02115*

**Abstract**

We present a greedy algorithm for supervised discretization using a metric defined on the space of partitions of a set of objects. This proposed technique is useful for preparing the data for classifiers that require nominal attributes. Experimental work on decision trees and naive Bayes classifiers confirm the efficacy of the proposed algorithm.

*Key words:* generalized entropy, metric, boundary points, naive Bayes classifiers

## 1 Introduction

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 ranging from fixed $k$-interval discretization [1], fuzzy discretization (see [2,3]), Shannon-entropy discretization due to Fayyad and Irani presented in [4,5], proportional

* Corresponding Author
*Email addresses:* rickb@cs.umb.edu (Richard Butterworth), dsim@cs.umb.edu (Dan A. Simovici), gsantos@mit.edu (Gustavo S. Santos).

$k$-interval discretization (see [6,7]), or techniques that are capable of dealing with highly dependent attributes (cf.[8]).

The discretization process can be described generically as follows. Let $B$ be a numerical attribute of a set of objects. The set of values of the components of these objects that correspond to the $B$ attribute is *the active domain of* $B$ and is denoted by $\mathsf{adom}(B)$.

To discretize $B$ we select a sequence of numbers $t_1 < t_2 < \cdots < t_\ell$ in $\mathsf{adom}(B)$. 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 $\hat{B}$-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 numbers $t_1, t_2, \ldots, t_\ell$ define the discretization process and they will be referred to as *class separators*.

We review briefly the terminology used in this paper. A *partition* of a non-empty set $S$ is a non-empty collection of non-empty subsets of $S$ indexed by a set $I$, $\pi = \{P_i \mid i \in I\}$ such that $\bigcup\{P_i \mid i \in I\} = S$, and $i, j \in I$, $i \neq j$ implies $P_i \cap P_j = \emptyset$. The sets $P_i$ are referred to as the *blocks of the partition* $\pi$. The set of partitions of $S$ is denoted by $\mathsf{PART}(S)$.

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. Recall that SQL computes the partition $\kappa_A$ using the `group by A` option of a `select` phrase.

There are two types of discretization [9]: *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 an attribute $B$ amounts to constructing a partition of the active domain $\mathsf{adom}(B)$ taking into account the partition $\kappa_A$ determined by the nominal class attribute $A$.

A partition $\pi = \{P_1, \ldots, P_k\}$ of a finite set $S$ generates a random variable:

$$X_\pi = \begin{pmatrix} 1 & 2 & \cdots & k \\ p_1 & p_2 & \cdots & p_k \end{pmatrix},$$

where $p_i = \frac{|P_i|}{|S|}$. This allows us to define Shannon's entropy of $\pi$ as the entropy of the random variable $X_\pi$, namely:

$$\mathcal{H}(\pi) = -\sum_{i=1}^{k} p_i \log_2 p_i.$$

For a subset $L$ of $S$ the *trace of the partition* $\pi$ on the set $L$ is the partition

$$\pi_L = \{P_i \cap L \mid 1 \leq i \leq k \text{ and } P_i \cap L \neq \emptyset\}.$$

Entropy measures the dispersion of values of a random variable. The maximum entropy for a $k$-valued random variable is obtained when $p_1 = \cdots = p_k = \frac{1}{k}$ and equals $\log_2 k$. Thus, the entropy of a partition $\pi_L$ serves to measure the scattering of the set $L$ across the blocks of $\pi$, that is, the impurity of the set $L$ relative to the partition $\pi$: the larger the entropy, the more $L$ is scattered among the blocks of $\pi$. If $\pi, \sigma$ are two partitions in $\mathsf{PART}(S)$, the average impurity of the blocks of $\sigma$ relative to $\pi$ is the *conditional entropy of* $\pi$ *relative to* $\sigma$:

$$\mathcal{H}(\pi|\sigma) = \sum_{j=1}^{m} \frac{|Q_j|}{|S|} \mathcal{H}(\pi_{Q_j}),$$

where $\sigma = \{Q_1, \ldots, Q_m\}$ and $\pi_{Q_j} = \{P_i \cap Q_j \mid P_i \in \pi \text{ and } P_i \cap Q_j \neq \emptyset\}$.

In [10] López de Màntaras proved that the function $d : \mathsf{PART}(S) \times \mathsf{PART}(S) \longrightarrow \mathbb{R}$ defined by: $d(\pi, \sigma) = \mathcal{H}(\pi|\sigma) + \mathcal{H}(\sigma|\pi)$, where $\mathcal{H}$ is Shannon's entropy is a metric on $\mathsf{PART}(S)$ (see [10]). Several authors have introduced generalizations of entropy (see [11–13]). The common nature of these generalizations have been highlighted by us in [14], where a unified axiomatization was introduced. Daróczy's $\beta$-entropy for a partition $\pi = \{P_1, \ldots, P_k\} \in \mathsf{PART}(S)$ is:

$$\mathcal{H}_\beta(\pi) = \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{i=1}^{k} \left(\frac{|P_i|}{|S|}\right)^\beta\right),$$

where $\beta$ is a positive number. It can be shown that $\lim_{\beta \to 1} \mathcal{H}_\beta(\pi)$ is Shannon's entropy.

For $\sigma, \pi \in \mathsf{PART}(S)$, where $\pi = \{P_1, \ldots, P_k\}$ and $\sigma = \{Q_1, \ldots, Q_m\}$, Daróczy's conditional $\beta$-entropy $\mathcal{H}_\beta(\pi|\sigma)$ is given by

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

Since

$$\mathcal{H}_\beta(\pi_{Q_j}) = \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{i=1}^{k} \left(\frac{|P_i \cup Q_j|}{|Q_j|}\right)^\beta\right)$$

we have

$$\mathcal{H}_\beta(\pi|\sigma) = \frac{1}{1-2^{1-\beta}} \sum_{j=1}^{m} \left(\frac{|Q_j|}{|S|}\right)^\beta \left(1 - \sum_{i=1}^{k} \left(\frac{|P_i \cup Q_j|}{|Q_j|}\right)^\beta\right),$$

which yields the useful equivalent expression:

$$\mathcal{H}_\beta(\pi|\sigma) = \frac{1}{(1-2^{1-\beta})|S|^\beta} \left(\sum_{j=1}^{m} |Q_j|^\beta - \sum_{i=1}^{k}\sum_{j=1}^{m} |P_i \cap Q_j|^\beta\right).$$

A related result obtained in [15] shows that the function $d_\beta : \mathsf{PART}(S) \times \mathsf{PART}(S) \longrightarrow \mathbb{R}$ given by:

$$\begin{aligned} d_\beta(\pi,\sigma) &= \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) \qquad (1)\\ &= \frac{1}{(1-2^{1-\beta})|S|^\beta} \left(\sum_{j=1}^{m} |Q_j|^\beta - \sum_{i=1}^{k}\sum_{j=1}^{m} |P_i \cap Q_j|^\beta + \right.\\ &\quad \left. \sum_{i=1}^{k} |P_i|^\beta - \sum_{i=1}^{k}\sum_{j=1}^{m} |P_i \cap Q_j|^\beta\right)\\ &= \frac{1}{(1-2^{1-\beta})|S|^\beta} \left(\sum_{i=1}^{n} |P_i|^\beta + \sum_{j=1}^{m} |Q_j|^\beta - 2 \cdot \sum_{i=1}^{k}\sum_{j=1}^{m} |P_i \cap Q_j|^\beta\right) \end{aligned}$$

is a metric. This distance was used in [15] to obtain small and accurate decision trees in an extension of López de Màntaras (see [10]) algorithm for building decision trees that makes use of Shannon's entropy.

For $\pi, \sigma \in \mathsf{PART}(S)$ we write $\pi \leq \sigma$ if each block of $\pi$ is included in a block of $\sigma$, or equivalently, if each block of $\sigma$ is an union of blocks of $\pi$. The partition $\sigma$ *covers* the partition $\pi$ (denoted by $\pi \prec \sigma$) if $\pi \leq \sigma$ and there is no partition $\theta \in \mathsf{PART}(S) - \{\pi, \sigma\}$ such that $\pi \leq \theta \leq \sigma$. This is equivalent to saying that $\sigma$ is obtained from $\pi$ by fusing together two blocks of $\pi$. If $\pi_1, \pi_2 \in \mathsf{PART}(S)$, then we denote by $\pi_1 \wedge \pi_2$ the partition whose blocks are all non-empty intersections of the form $K \cap H$, where $K \in \pi_1$ and $H \in \pi_2$. The least partition of $\mathsf{PART}(S)$ is the partition $\iota_S = \{\{x\} \mid x \in S\}$ whose blocks are the singletons of $S$; the largest partition of $\mathsf{PART}(S)$ is the one-block partition $\omega_S = \{S\}$.

The generalized conditional entropy is dually monotonic in its first argument and monotonic in its second, that is $\pi \leq \pi'$ implies $\mathcal{H}_\beta(\pi|\sigma) \geq \mathcal{H}_\beta(\pi'|\sigma)$ and $\sigma \leq \sigma'$ implies $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi|\sigma')$, as we have shown in [15].

Partitions of active attribute domains induce partitions on the set of objects. Namely, the partition of the set of objects $S$ that corresponds to a partition $\pi$ of $\mathsf{adom}(B)$, where $B$ is a numerical attribute, is denoted by $\pi_*$. A block of

$\pi_*$ consists of all objects whose $B$-components belong to the same block of $\pi$. For the special case when $\pi = \iota_{\mathsf{adom}(B)}$ observe that $\pi_* = \kappa_B$.

Let $\mathsf{T} = (t_1, \ldots, t_\ell)$ be the sequence of class separators of the active domain of an attribute $B$, where $t_1 < t_2 < \cdots < t_\ell$. This set of cutpoints creates a partition $\pi_B^{\mathsf{T}} = \{Q_0, \ldots, Q_\ell\}$ of $\mathsf{adom}(B)$, where $Q_i = \{b \in \mathsf{adom}(B) \mid t_i \leq b < t_{i+1}\}$ for $0 \leq i \leq \ell$, where $t_0 = -\infty$ and $t_{\ell+1} = +\infty$.

It is immediate that for two sets of cutpoints $\mathsf{T}, \mathsf{T}'$ we have $\pi_B^{\mathsf{T} \cup \mathsf{T}'} = \pi_B^{\mathsf{T}} \wedge \pi_B^{\mathsf{T}'}$. If the sequence $\mathsf{T}$ consists of a single cutpoint $t$ we shall denote $\pi_B^{\mathsf{T}}$ simply by $\pi_B^t$. The discretization process consists of replacing each value that falls in the block $Q_i$ of $\pi_B^{\mathsf{T}}$ by $i$ for $0 \leq i \leq \ell$.

Suppose that the list of objects sorted on the values of a numerical attribute $B$ is $o_1, \ldots, o_n$ and let $o_1[B], \ldots, o_n[B]$ be the sequence of $B$-components of those objects, where $o_1[B] \leq o_2[B] \leq \cdots \leq o_n[B]$. For a nominal attribute $A$ define the partition $\pi_{B,A}$ of $\mathsf{adom}(B)$ as follows. A block of $\pi_{B,A}$ consists of a maximal subsequence $o_i[B], \ldots, o_l[B]$ of the previous sequence such that every object $o_i, \ldots, o_l$ of this subsequence belongs to the *same block* $K$ of the partition $\kappa_A$. If $x \in \mathsf{adom}(B)$, we shall denote the block of $\pi_{B,A}$ that contains $x$ by $\langle x \rangle$. The *boundary points* of the partition $\pi_{B,A}$ are the least and the largest elements of each of the blocks of the partition $\pi_{B,A}$. The least and the largest elements of $\langle x \rangle$ are denoted by $x^\downarrow$ and $x^\uparrow$, respectively. It is clear that $\pi_{B,A*} \leq \kappa_A$ for any attribute $B$.

**Example 1.1** Let $o_1, \ldots, o_9$ be a collection of nine objects such that the sequence $o_1[B], \ldots, o_9[B]$ is sorted in increasing order of the value of the $B$-components:

| | $\cdots$ | $B$ | $\cdots$ | $A$ |
|---|---|---|---|---|
| $o_1$ | $\cdots$ | 95.2 | $\cdots$ | Y |
| $o_2$ | $\cdots$ | 110.1 | $\cdots$ | N |
| $o_3$ | $\cdots$ | 120.0 | $\cdots$ | Y |
| $o_4$ | $\cdots$ | 125.5 | $\cdots$ | Y |
| $o_5$ | $\cdots$ | 130.1 | $\cdots$ | N |
| $o_6$ | $\cdots$ | 140.0 | $\cdots$ | N |
| $o_7$ | $\cdots$ | 140.5 | $\cdots$ | Y |
| $o_8$ | $\cdots$ | 168.2 | $\cdots$ | Y |
| $o_9$ | $\cdots$ | 190.5 | $\cdots$ | Y |

The partition $\kappa_A$ has two blocks corresponding to the values 'Y' and 'N' and

is given by:

$$\kappa_A = \{\{o_1, o_3, o_4, o_7, o_8, o_9\}, \{o_2, o_5, o_6\}\}.$$

The partition $\pi_{B,A*}$ is:

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

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 [4] showed that to obtain the least value of the Shannon's conditional entropy $\mathcal{H}(\pi_A|\pi^{\mathsf{T}}_{B*})$ the cutpoints $t$ of $\mathsf{T}$ must be chosen among the boundary points of the the partition $\pi_{B,A}$. This is a powerful result that limits drastically the number of possible cut points and improves the tractability of the discretization.

We present two new basic ideas: a generalization of Fayyad-Irani discretization techniques that relies on a metric on partitions defined by Daróczy's generalized entropy, and a new geometric criterion for halting the discretization process. With an appropriate choice of the parameters of the discretization process the resulting decision trees are smaller, have fewer leaves, and display higher levels of accuracy as verified by stratified cross-validation; similarly, naive Bayes classifiers applied to data discretized by our algorithm yield smaller error rates.

Our main results show that the same choice of cutpoints must be made for a broader class of impurity measures, namely the impurity measures related to generalized conditional entropy. Moreover, when the purity of the partition $\pi_A$ is replaced as a discretization criterion by the minimality of the entropic distance between the partitions $\pi_A$ and $\pi^{\mathsf{T}}_{B,*}$ (introduced in [15]) the same method for selecting the cutpoint can be applied. This is a generalization of the approach proposed by Cerquides and López de Màntaras in [16]).

## 2 A Generalization of Fayyad's Result

We are concerned with supervised discretization, that is, with discretization of attributes that takes into account the classes where the objects belong. Suppose that the class of objects is determined by the nominal attribute $A$ and we need to discretize a numerical attribute $B$. The discretization of $B$ aims to construct a set $\mathsf{T}$ of cutpoints of $\mathsf{adom}(B)$ such that the blocks of $\kappa_A$ are as pure as possible relative to the partition $\pi^{\mathsf{T}}_{B*}$, that is, the conditional entropy $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})$ is minimal.

The following theorem extends a result of Fayyad (Theorem 5.4.1 of [4]):

**Theorem 2.1** *Let $S$ be a collection of objects where the class of an object is determined by the attribute $A$ and let $\beta \in (1, 2]$. If $\mathsf{T}$ is a set of cutpoints such that the conditional entropy $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})$ is minimal among the set of cutpoints with the same number of elements, then $\mathsf{T}$ consists of boundary points of the partition $\pi_{B,A}$ of $\mathsf{adom}(B)$.*

**Proof.** See Appendix A.1. □

The next theorem is a companion to Fayyad's result and makes use of the same hypothesis as Theorem 2.1.

**Theorem 2.2** *Let $\beta$ be a number, $\beta \in (1, 2]$. If $\mathsf{T}$ is a set of cutpoints of $\mathsf{adom}(B)$ such that the distance $d_\beta(\kappa_A, \pi^{\mathsf{T}}_{B*})$ is minimal among the set of cutpoints with the same number of elements, then $\mathsf{T}$ consists of boundary points of the partition $\pi_{B,A}$ of $\mathsf{adom}(B)$.*

**Proof.** The argument for this statement is given in Appendix A.2. □

This result will play a key role in the algorithm that we propose in this paper. To discretize $\mathsf{adom}(B)$ we shall seek a set of cutpoints $\mathsf{T}$ such that $d_\beta(\kappa_A, \pi^{\mathsf{T}}_{B*}) = \mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*}) + \mathcal{H}_\beta(\pi^{\mathsf{T}}_{B*}|\kappa_A)$ is minimal. In other words, we shall seek a set of cutpoints such that the partition $\pi^{\mathsf{T}}_{B*}$ induced on the set of objects $S$ is as close as possible to the target partition $\kappa_A$.

Initially, before adding cutpoints, we have $\mathsf{T} = \emptyset$, $\pi^{\mathsf{T}}_{B*} = \omega_S = \{S\}$, and therefore $\mathcal{H}_\beta(\kappa_A|\omega_S) = \mathcal{H}_\beta(\kappa_A)$. Observe that when the set $\mathsf{T}$ grows the entropy $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})$ decreases. Note that the use of conditional entropy $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})$ tends to favor large cutpoint sets for which the partition $\pi^{\mathsf{T}}_{B*}$ is small in the partial ordered set $(\mathsf{PART}(\mathsf{T}), \leq)$. In the extreme case, every point would be a cutpoint, a situation that is clearly unacceptable. Fayyad-Irani technique halts the discretization process using the principle of minimum description. We adopt another technique that has the advantage of being geometrically intuitive and produces very good experimental results.

Using the distance $d_\beta(\kappa_A, \pi^{\mathsf{T}}_{B*}) = \mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*}) + \mathcal{H}_\beta(\pi^{\mathsf{T}}_{B*}|\kappa_A)$ the decrease of $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})$ when the set of cutpoints grows is balanced by the increase in $\mathcal{H}_\beta(\pi^{\mathsf{T}}_{B*}|\kappa_A)$. Note that initially we have $\mathcal{H}_\beta(\omega_S|\kappa_A) = 0$. The discretization process can thus be halted when the distance $d_\beta(\kappa_A, \pi^{\mathsf{T}}_{B*})$ stops decreasing. Thus, we retain as a set of cutpoints for discretization the set $\mathsf{T}$ that determines the closest partition to the class partition $\kappa_A$. As a result, we obtain good discretizations (as evaluated through the results of various classifiers that use the discretize data) with relatively small cutpoint sets.

## 3 Discretization Algorithm and Experimental Results

The greedy algorithm shown below is used for discretizing an attribute $B$. It makes successive passes over the table and, at each pass it adds a new cutpoint chosen among the boundary points of $\pi_{B,A}$.

**Input:** A table $S$, a class attribute $A$,
and a real-valued attribute $B$.
**Output:** A discretized attribute $B$.
**Method:** sort table $S$ on attribute $B$;
compute the set $\mathsf{BP}$ of boundary points of partition $\pi_{B,A}$;
$\mathsf{T} = \emptyset$; $d = \infty$;
while $\mathsf{BP} \neq \emptyset$ do
    let $t = \arg\min_{t \in \mathsf{BP}} d_\beta(\kappa_A, \pi_{B*}^{\mathsf{T}\cup\{t\}})$;
    if $d \geq d_\beta(\kappa_A, \pi_{B*}^{\mathsf{T}\cup\{t\}})$ then
        begin
            $\mathsf{T} = \mathsf{T} \cup \{t\}$;
            $\mathsf{BP} = \mathsf{BP} - \{t\}$;
            $d = d_\beta(\kappa_A, \pi_{B*}^{\mathsf{T}})$
        end
    else
        exit while loop;
end while
for $\pi_{B*}^{\mathsf{T}} = \{Q_0, \ldots, Q_\ell\}$ replace
every attribute in $Q_i$ by $i$ for $0 \leq i \leq \ell$.

The while loop is running for as long as there exist candidate boundary points and it is possible to find a new cutpoint $t$ such that the distance $d_\beta(\kappa_A, \pi_{B*}^{\mathsf{T}\cup\{t\}})$ is less than the previous distance $d_\beta(\kappa_A, \pi_{B*}^{\mathsf{T}})$. An experiment performed on a synthetic database shows that a substantial amount of time (about 78% of the total time) is spent on decreasing the distance by the last 1% (see Figure 1). Therefore, in practice we run a search for a new cutpoint only if $|d - d_\beta(\kappa_A, \pi_{B*}^{\mathsf{T}\cup\{t\}})| > 0.01d$.

To form an idea on the evolution of the distance between $\kappa_A$ and the partition of objects determined by the cutpoints $\pi_{B*}^{\mathsf{T}}$ let $t \in \mathsf{BP}$ be a new cutpoint added to the set $\mathsf{T}$. It is clear that the partition $\pi_{B*}^{\mathsf{T}}$ covers the partition $\pi_{B*}^{\mathsf{T}\cup\{t\}}$ because $\pi_{B*}^{\mathsf{T}\cup\{t\}}$ is obtained by splitting a block of $\pi_{B*}^{\mathsf{T}}$. Without loss of generality we assume that the blocks $Q_{m-1}$ and $Q_m$ of $\pi_{B*}^{\mathsf{T}\cup\{t\}}$ result from the split of the block $Q_{m-1} \cup Q_m$ of $\pi_{B*}^{\mathsf{T}}$:

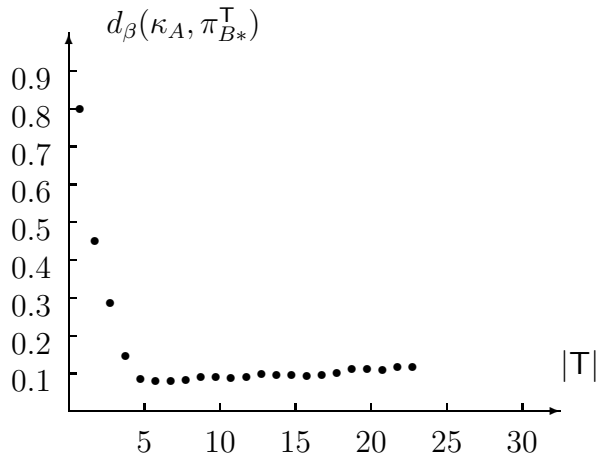


Fig. 1. Variation of Distance with the Size of the Set of Cutpoints

$$\begin{aligned}\kappa_A &= \{P_1, \ldots, P_n\},\\ \pi^{\mathsf{T}}_{B*} &= \{Q_1, \ldots, Q_{m-2}, Q_{m-1} \cup Q_m\}\\ \pi^{\mathsf{T}\cup\{t\}}_{B*} &= \{Q_1, \ldots, Q_{m-2}, Q_{m-1}, Q_m\}.\end{aligned}$$

Since $\beta > 1$, by Equality (1), we have $d_\beta(\kappa_A, \pi^{\mathsf{T}\cup\{t\}}_{B*}) < d_\beta(\kappa_A, \pi^{\mathsf{T}}_{B*})$ if and only if

$$\begin{aligned}&\textstyle\sum_{i=1}^{n} |P_i|^\beta + \sum_{j=1}^{m} |Q_j|^\beta - 2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} |P_1 \cap Q_j|^\beta <\\ &\textstyle\sum_{i=1}^{n} |P_i|^\beta + \sum_{j=1}^{m-2} |Q_j|^\beta + |Q_{m-1} \cup Q_m|^\beta\\ &\textstyle-2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{m-2} |P_i \cap Q_j|^\beta - 2 \cdot \sum_{i=1}^{n} |P_i \cap (Q_{m-1} \cup Q_m)|^\beta,\end{aligned}$$

which is equivalent to:

$$\begin{aligned}&|Q_{m-1}|^\beta + |Q_m|^\beta - 2 \cdot \textstyle\sum_{i=1}^{n} |P_i \cap Q_{m-1}|^\beta - 2 \cdot \sum_{i=1}^{n} |P_i \cap Q_m|^\beta <\\ &|Q_{m-1} \cup Q_m|^\beta - 2 \cdot \textstyle\sum_{i=1}^{n} \left(|P_i \cap Q_{m-1}| + |P_i \cap Q_m|\right)^\beta.\end{aligned}$$

Suppose that $Q_{m-1} \cup Q_m$ is intersected by only by $P_1$ and $P_2$ and that $\beta = 2$. Then, the previous inequality that describes the condition under which a decrease of $d_\beta(\kappa_A, d^{\mathsf{T}}_*)$ can be obtained becomes:

$$(|P_1 \cap Q_{m-1}| - |P_2 \cap Q_{m-1}|)(|P_1 \cap Q_m| - |P_2 \cap Q_m|) < 0, \tag{2}$$

and so, the distance may be decreased by splitting a block $Q_{m-1} \cup Q_m$ into $Q_{m-1}$ and $Q_m$, only when the distribution of the fragments of the blocks $P_1$ and $P_2$ in the prospective blocks $Q_{m-1}$ and $Q_m$ satisfies condition (2). If the block $Q_{m-1} \cup Q_m$ of the partition $\pi^{\mathsf{T}}_{B*}$ contains a unique boundary point, then choosing that boundary point as a cutpoint will decrease the distance. Indeed, in this case we have $|P_1 \cap Q_{m-1}| > 0$, $|P_1 \cap Q_m| = 0$, and $|P_2 \cap Q_{m-1}| = 0$, $|P_2 \cap Q_m| > 0$, which guarantees that condition (2) is satisfied.

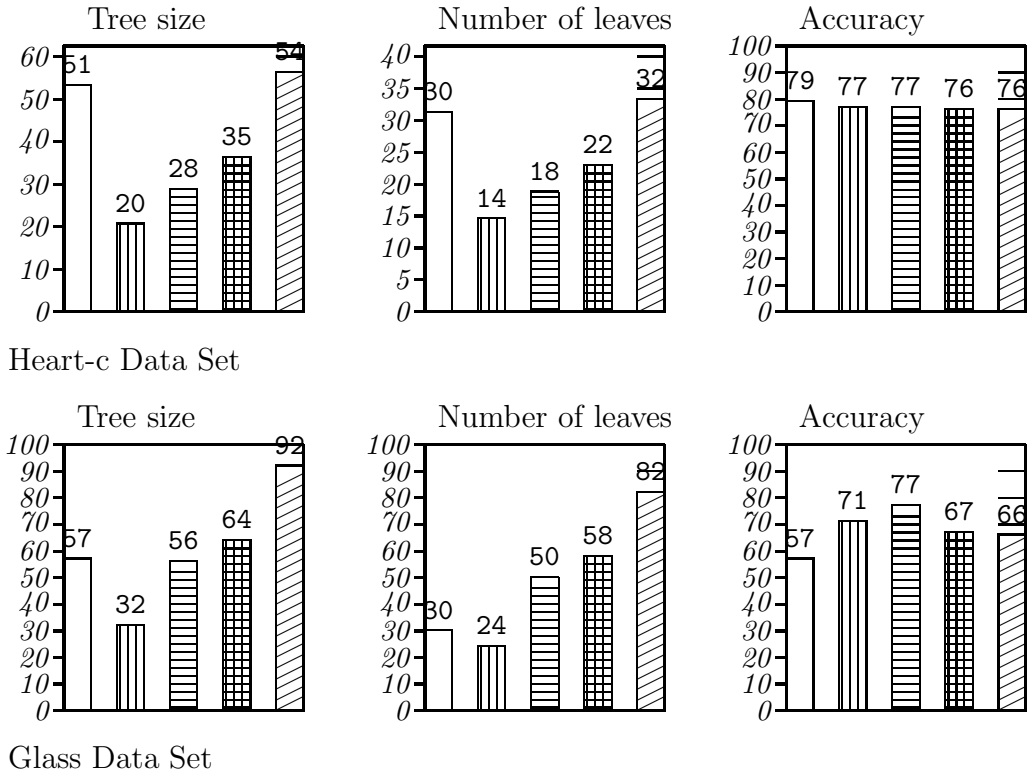


Fig. 2. Experimental Results for the Heart-c and Glass Data Sets

We tested our discretization algorithm on several machine learning data sets from UCI data sets [17] 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 [9]. The size, number of leaves and accuracy of the trees are described in Table 1, where trees built using the Fayyad-Irani discretization method of J48 are designated as "standard".

It is clear that 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 as shown by the experiments on the glass data set. (see Figure 2).

Our supervised discretization algorithm that discretizes each attribute $B$ based on the relationship between the partition $\pi_B$ and $\pi_A$ (where $A$ is the attribute that specifies the class of the objects). Thus, the discretization process of an

| Database | Experimental Results | | | |
|---|---|---|---|---|
| | Discretization method | Size | Number of leaves | Accuracy (stratified cross-validation) |
| heart-c | *standard* | *51* | *30* | *79.20* |
| | $\beta = 1.5$ | 20 | 14 | 77.36 |
| | $\beta = 1.8$ | 28 | 18 | 77.36 |
| | $\beta = 1.9$ | 35 | 22 | 76.01 |
| | $\beta = 2.0$ | 54 | 32 | 76.01 |
| glass | *standard* | *57* | *30* | *57.28* |
| | $\beta = 1.5$ | 32 | 24 | 71.02 |
| | $\beta = 1.8$ | 56 | 50 | 77.10 |
| | $\beta = 1.9$ | 64 | 58 | 67.57 |
| | $\beta = 2.0$ | 92 | 82 | 66.35 |
| ionosphere | *standard* | *35* | *18* | *90.88* |
| | $\beta = 1.5$ | 15 | 8 | 95.44 |
| | $\beta = 1.8$ | 19 | 12 | 88.31 |
| | $\beta = 1.9$ | 15 | 10 | 90.02 |
| | $\beta = 2.0$ | 15 | 10 | 90.02 |
| iris | *standard* | *9* | *5* | *95.33* |
| | $\beta = 1.5$ | 7 | 5 | 96 |
| | $\beta = 1.8$ | 7 | 5 | 96 |
| | $\beta = 1.9$ | 7 | 5 | 96 |
| | $\beta = 2.0$ | 7 | 5 | 96 |
| diabetes | *standard* | *43* | *22* | *74.08* |
| | $\beta = 1.8$ | 5 | 3 | 75.78 |
| | $\beta = 1.9$ | 7 | 4 | 75.39 |
| | $\beta = 2.0$ | 14 | 10 | 76.30 |

Table 1
Comparative Experimental Results for Decision Trees

| Discretization Method | Diabetes | Glass | Ionosphere | Iris |
|---|---|---|---|---|
| $\beta = 1.5$ | 34.9 | 25.2 | 4.8 | 2.7 |
| $\beta = 1.8$ | 24.2 | 22.4 | 8.3 | 4 |
| $\beta = 1.9$ | 24.9 | 23.4 | 8.5 | 4 |
| $\beta = 2.0$ | 25.4 | 24.3 | 9.1 | 4.7 |
| weighted proportional | 25.5 | 38.4 | 10.3 | 6.9 |
| proportional | 26.3 | 33.6 | 10.4 | 7.5 |

Table 2
Error Rate for Naive Bayes Classifiers

attribute is carried out independently of similar processes performed on other attributes. As a result, our algorithm is particularly efficient for naive Bayes classifiers, which rely on the essential assumption of attribute independence. The error rates of Naive Bayes Classifiers obtained for different discretization methods are shown in Table 2.

We applied the proposed discretization method to a data set [18] that is obtained from the use of microarray technology and is used in the diagnostic of differential diagnosis of small round-blue cell tumors (SRCBCF) of childhood: neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt lymphoma (BL), and the Ewing family of tumors (EWS). The training data include 63 samples (23 EWS, 20 RMS, 12 NB, and 8 BL) with 6567 genes used in the model; the test data include 25 samples (6 EWS, 5 RMS, 6NB, 3BL, and 5 "noise samples" originating in other tissues). An investigation of this set of data was carried out in [19] using fuzzy logic.

For each gene $G$ involved we computed the distance $d_\beta(\pi_G, \pi_D)$ (where $D$ is the diagnosis attribute). The discretization process involved 30 genes $G$ having the least 30 values for $d_\beta(\pi_G, \pi_D)$. We applied discretization to the training set for several values of $\beta$ and stopped the discretization algorithm after the first two cutting points were detected. Then, in each case, a naive Bayes classifier was constructed using the WEKA package [9]. The results are shown in Table 3. The results suggest that the optimal value of $\beta$ for this data set is 1.4.

## 4 Conclusions and Open Problems

The use of the metric space of partitions of the data set in discretization is helpful in preparing the data for classifiers. With an appropriate choice of the

| Discretization Method | Accuracy Rate on Test Set | Misclassified | |
|---|---|---|---|
| | | "noise" cases | regular cases |
| $\beta = 1.3$ | 76% | 5 | 1 |
| $\beta = 1.35$ | 60% | 4 | 6 |
| $\beta = 1.4$ | 84% | 3 | 1 |
| $\beta = 1.5$ | 80% | 2 | 3 |

Table 3
Accuracy Rate on Test Set on Khan's Data

parameter $\beta$ that defines the metric used in discretization, standard classifiers such as C4.5 or J48 generate smaller decision trees with comparable or better levels of accuracy when applied to data discretized with our technique.

An important open issue is determining characteristics of data sets that will inform the choice of an optimal value for the $\beta$ parameter.

Also, investigating metric discretization for data with missing values seems to present particular challenges that we intend to consider in our future work.

## 5 Acknowledgement

The authors would like to express their gratitude to the reviewers whose observations improved the readability of this paper.

## A Proofs of Theorems

### *A.1 Proof of Theorem 2.1*

The proof is by induction on the number of cutpoints $\ell = |\mathsf{T}|$. If $\ell = 0$, the statement is immediate since in this case $\pi^{\mathsf{T}}_{B*}$ is the one-class partition $\omega_S$ of the set of objects $S$.

Suppose that the statement holds for set of cutpoints that contain $\ell$ elements and let $\mathsf{Z} = \mathsf{T} \cup \{t\}$, where $\mathsf{T} = \{t_1, \ldots, t_\ell\}$ is a set of cutpoints that is a subset of the set of boundary points of $\pi_{B,A}$, $|\mathsf{T}| = \ell$ and $t \notin T$.

Let $\kappa_A = \{P_1, \ldots, P_k\}$ and $\pi^{\mathsf{T}}_{B*} = \{Q_0, \ldots, Q_\ell\}$, where $\kappa_A, \pi^{\mathsf{T}}_{B*} \in \mathsf{PART}(S)$.

The conditional entropy $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})$ is given by:

$$\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*}) = \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(\sum_{j=0}^{\ell}|Q_j|^\beta - \sum_{i=1}^{k}\sum_{j=0}^{\ell}|P_i\cap Q_j|^\beta\right).$$

Suppose that the new cut point $t$ is placed between $t_{h-1}$ and $t_h$. Then, the partition $\pi^{\mathsf{Z}}_{B*}$ is obtained from $\pi^{\mathsf{T}}_{B*}$ by splitting $Q_h$ in $Q'_h$ and $Q''_h$. Also, $t$ is located between two cutpoints $t^\downarrow$ and $t^\uparrow$ of the partition $\pi_{B,A}$. Since $\pi_{B,A*} \leq \pi_A$ the set of objects whose $B$-component is included in the interval $\langle t\rangle = [t^\downarrow, t^\uparrow]$ is a subset of a block $P_g$ of the partition $\kappa_A$.

The variation of the entropy caused by the introduction of the split in $Q_h$ is given by:

$$\begin{aligned}
&\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{Z}}_{B*}) - \mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})\\
&= \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(\sum_{j=0,j\neq h}^{\ell}|Q_j|^\beta - \sum_{i=1}^{k}\sum_{j=0,j\neq h}^{\ell}|P_i\cap Q_j|^\beta\right)\\
&\quad + \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(|Q'_h|^\beta + |Q''_h|^\beta - \sum_{i=1}^{k}|P_i\cap Q'_h|^\beta + \sum_{i=1}^{k}|P_i\cap Q''_h|^\beta -\right)\\
&\quad - \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(\sum_{j=0}^{\ell}|Q_j|^\beta - \sum_{i=1}^{k}\sum_{j=0}^{\ell}|P_i\cap Q_j|^\beta\right).
\end{aligned}$$

Since the partition $\pi^{\mathsf{T}}_{B*}$ is such that $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})$ achieves a local minimum, it follows that the difference $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{Z}}_{B*}) - \mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})$ needs to have a local minimum in order for $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{Z}}_{B*})$ to achieve a local minimum.

The number of objects in the sets $P_i\cap Q_j$ for $i\neq g$ and $j\neq h$ is unaffected by the split of $Q_h$ since $\langle t\rangle \subseteq P_g$. There is a constant $K$ (independent of $t$) such that the variation in entropy can be written as

$$\begin{aligned}
&\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{Z}}_{B*}) - \mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})\\
&= \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(|Q'_h|^\beta + |Q''_h|^\beta\right.\\
&\quad \left. -K - |P_g\cap Q'_h|^\beta - |P_g\cap Q''_h|^\beta\right).
\end{aligned}$$

Denote $n = |\langle t\rangle|$, and let $\mu$ be the number of objects whose $B$-component is in $(t^\downarrow, t]$. Then, the number of objects whose $B$-component is in $(t, t^\uparrow]$ is $n-\mu$. Let $a, b$ be the numbers of objects in $Q'_h$ and $Q''_h$ whose $B$-component is less than $t^\downarrow$ and $t^\uparrow$, respectively. With these notations we can write

$$\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{Z}}_{B*}) - \mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*}) = \frac{1}{(1-2^{1-\beta})|S|^\beta}\Big((a+\mu)^\beta + (b+n-\mu)^\beta \\ -K-\mu^\beta-(n-\mu)^\beta\Big),$$

If we regard $\mu$ as a continuous variable varying in the interval $[0, n]$ we need to examine the variation of the real-valued function

$$F(\mu) = \frac{1}{(1-2^{1-\beta})|S|^\beta}\Big((a+\mu)^\beta + (b+n-\mu)^\beta - K - \mu^\beta - (n-\mu)^\beta\Big),$$

on the interval $[0, n]$. The second derivative of this function is:

$$F''(\mu) = \frac{\beta(\beta-1)}{(1-2^{1-\beta})|A|^\beta}\Big((a+\mu)^{\beta-2} + (b+n-\mu)^{\beta-2} \\ -\mu^{\beta-2} - (n-\mu)^{\beta-2}\Big).$$

Since $\beta > 1$ we have $\frac{\beta(\beta-1)}{1-2^{1-\beta}} > 0$. Also, for $1 \leq \beta < 2$ we have both $\mu^{\beta-2} - (a+\mu)^{\beta-2} > 0$ and $(n-\mu)^{\beta-2} - (b+n-\mu)^{\beta-2} > 0$, which imply that the second derivative is negative on $[0, n]$. This proves that the minimum of this function is attained either for $\mu = 0$ or for $\mu = n$, that is, in one of the $\pi_{B,A}$-boundary points.

The case $\beta = 2$ is immediate since in this situation $F$ is a linear function of $\mu$. $\square$

### *A.2 Proof of Theorem 2.2*

As before, the argument is by induction on $|\mathsf{T}|$ and the base case $|\mathsf{T}| = 0$ is vacuous. Suppose that the statement is true for $|\mathsf{T}| = \ell$, so $\mathsf{T}$ consists of boundary points of the partition $\pi_{B,A}$.

The conditional entropy $\mathcal{H}_\beta(\pi^{\mathsf{T}}_{B*}|\kappa_A)$ is given by

$$\mathcal{H}_\beta(\pi^{\mathsf{T}}_{B*}|\kappa_A) = \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(\sum_{i=1}^{k}|P_i|^\beta - \sum_{i=1}^{k}\sum_{j=0}^{\ell}|P_i \cap Q_j|^\beta\right).$$

If we add a new cutpoint $t$ between the boundary points $t_{h-1}$ and $t_h$ to obtain the new set of cutpoints $\mathsf{Z} = \mathsf{T} \cup \{t\}$, the new value of the conditional entropy is:

$$\mathcal{H}_\beta(\pi^{\mathsf{Z}}_{B*}|\kappa_A) = \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(\sum_{i=1}^{k}|P_i|^\beta - \sum_{i=1}^{k}\sum_{j=0,j\neq h}^{\ell}|P_i\cap Q_j|^\beta - \sum_{i=1}^{k}|P_i\cap Q'_h|^\beta - \sum_{i=1}^{k}|P_i\cap Q''_h|^\beta\right).$$

Thus, we have:

$$\mathcal{H}_\beta(\pi^{\mathsf{Z}}_{B*}|\kappa_A) - \mathcal{H}_\beta(\pi^{\mathsf{T}}_{B*}|\kappa_A) = \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(\sum_{i=1}^{k}|P_i\cap Q'_h|^\beta + \sum_{i=1}^{k}|P_i\cap Q''_h|^\beta + \sum_{i=1}^{k}|P_i\cap Q_h|^\beta\right).$$

Since $\langle t\rangle \subseteq P_g$ only the intersections that contain $P_g$ depend on the position of the new cutpoint $t$. Therefore, the variation of the conditional entropy can be written as

$$\begin{aligned}&\mathcal{H}_\beta(\pi^{\mathsf{Z}}_{B*}|\kappa_A) - \mathcal{H}_\beta(\pi^{\mathsf{T}}_{B*}|\kappa_A)\\ &= \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(H + |P_g\cap Q_h|^\beta - |P_g\cap Q'_h|^\beta - |P_g\cap Q''_h|^\beta\right),\end{aligned}$$

where $H$ is a constant that does not depend on $t$. Using the notation previously introduced we have

$$\begin{aligned}&\mathcal{H}_\beta(\pi^{\mathsf{Z}}_{B*}|\kappa_A) - \mathcal{H}_\beta(\pi^{\mathsf{T}}_{B*}|\kappa_A)\\ &= \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(H + n^\beta - \mu^\beta - (n-\mu)^\beta\right).\end{aligned}$$

The second derivative of the real-valued function $G$ defined by:

$$G(\mu) = \frac{1}{(1-2^{1-\beta})|S|^\beta}\left(H + n^\beta - \mu^\beta - (n-\mu)^\beta -\right)$$

for $\mu \in (0, n]$ is

$$G''(\mu) = -\frac{\beta(\beta-1)}{(1-2^{1-\beta})|S|^\beta}\left(\mu_{\beta-2} + (n-\mu)^{\beta-2}\right)$$

and is clearly negative.

The variation of the distance $d_\beta(\kappa_A, \pi^{\mathsf{Z}}_{B*}) - d_\beta(\kappa_A, \pi^{\mathsf{T}}_{B*})$ is the sum of the variations of the entropies $\mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{Z}}_{B*}) - \mathcal{H}_\beta(\kappa_A|\pi^{\mathsf{T}}_{B*})$ and $\mathcal{H}_\beta(\pi^{\mathsf{Z}}_{B*}|\kappa_A) - \mathcal{H}_\beta(\pi^{\mathsf{T}}_{B*}|\kappa_A)$. With the above notation, this variation equals $F(\mu) + G(\mu)$, where $F$ is the function introduced in the proof of Theorem 2.1. Since $F''(\mu) + G''(\mu) < 0$,

the minimum value of the distance can be attained only when $t$ coincides with either $t^{\downarrow}$ or with $t^{\uparrow}$. $\square$

## References


[1] J. Dougherty, R. Kohavi, M. Sahami, Supervised and unsupervised discretization of continuous features, in: Proc. of the 12th International Conference on Machine Learning, 1995, pp. 194–202.

[2] I. Kononenko, Naive bayes classifier and continuous attributes, Informatica 16 (1992) 1–8.

[3] I. Kononenko, Inductive and Bayesian learning in medical diagnosis, Applied Artificial Intelligence 7 (1993) 317–337.

[4] U. M. Fayyad, On the induction of decision trees for multiple concept learning, Ph.D. thesis, University of Michigan (1991).

[5] U. M. Fayyad, K. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: Proc. of the 12th International Joint Conference on Artificial Intelligence, 1993, pp. 1022–1027.

[6] Y. Yang, G. I. Webb, Proportional $k$-interval discretization for naive-Bayes classifiers, in: Proc. of the 12th European Conference on Machine Learning, 2001, pp. 564–575.

[7] Y. Yang, G. I. Webb, Weighted proportional $k$-interval discretization for naive-Bayes classifiers, in: Proc. of the PAKDD, 2003.

[8] M. Robnik, I. Kononenko, Discretization of continuous attributes using relieff, in: Proc. of ERK-95, 1995, pp. 149–152.

[9] I. H. Witten, E. Frank, Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, 2000.

[10] R. L. de Màntaras, A distance-based attribute selection measure for decision tree induction, Machine Learning 6 (1991) 81–92.

[11] P. A. Devijer, Entropie quadratique et reconnaisance des formes, in: Computer Oriented Learning Processes, Proceedings of the NATO Advanced Study Institute, Château de Bonas, France, 1974, pp. 257–278.

[12] Z. Daróczy, Generalized information functions, Information and Control 16 (1970) 36–51.

[13] J. H. Havrda, F. Charvat, Quantification methods of classification processes: Concepts of structural $\alpha$-entropy, Kybernetica 3 (1967) 30–35.

[14] D. A. Simovici, S. Jaroszewicz, An axiomatization of partition entropy, IEEE Transactions on Information Theory 48 (2002) 2138–2142.

[15] D. Simovici, S. Jaroszewicz, Generalized conditional entropy and decision trees, in: Extraction et Gestion des connaissances - EGC 2003, Lavoisier, Paris, 2003, pp. 363–380.

[16] J. Cerquides, R. L. de Màntaras, Proposal and empirical comparison of a parallelizable distance-based discretization method, in: Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'97), Newport Beach,CA, 1997, pp. 139–142.

[17] C. L. Blake, 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.

[18] J. Khan, J. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westerman, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, P. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine 7 (2001) 673–679.

[19] L. Ohno-Machado, S. A. Vinterbo, G. Webber, Classification of gene expression data using fuzzy logic, Journal of Intelligent and Fuzzy Systems 12 (2002) 19–24.