A greedy algorithm for supervised discretization

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

a Department of Computer Science, University of Massachusetts at Boston, Boston, MA 02125, USA
b Decision Systems Group, Division of Health Sciences and Technology, Harvard and MIT, Brigham and Women’s Hospital, 75 Francis Street, Boston, MA 02115, USA

Received 11 June 2004
Available online 27 August 2004

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.

© 2004 Elsevier Inc. All rights reserved.

Keywords: Generalized entropy; Metric; Boundary points; Naïve 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 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 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 adom(B).

To discretize B we select a sequence of numbers \( t_1 < t_2 < \cdots < t_{\ell} \) in adom(B). Next, the attribute B is replaced by the nominal attribute \(^B\) that has \( \ell + 1 \) distinct values in its active domain \( \{k_0, k_1, \ldots, k_{\ell}\} \). Each B-component b of an object o is replaced by the discretized \(^B\)-component defined by

\[
\begin{align*}
    k &= \begin{cases} 
        k_0 & \text{if } b \leq t_1, \\
        k_i & \text{if } t_i < b \leq t_{i+1} \text{ for } 1 \leq i \leq \ell - 1, \\
        k_\ell & \text{if } t_\ell < b.
    \end{cases}
\end{align*}
\]

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, \( \pi = \{P_i | i \in I\} \) such that \( \bigcup \{P_i | i \in I\} = S \), 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 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...
Entropy measures the dispersion of values of a random variable. The average impurity of the blocks of a partition \( \pi \) is defined 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 \( \log k \). 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 \( \text{PART}(S) \), 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 \( \pi_{Q_j} = \{P_i \cap Q_j \mid P_i \in \pi \text{ and } P_i \cap Q_j \neq \emptyset \} \).

López de Mántaras [10] proved that the function \( d : \text{PART}(S) \times \text{PART}(S) \rightarrow \mathbb{R} \) defined by:

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

where \( \mathcal{H} \) is Shannon’s entropy, is a metric on \( \text{PART}(S) \) (see [10]). Several authors have introduced generalizations of entropy (see [11-13]). The common nature of these generalizations has been highlighted by us in [14], where a unified axiomatization was introduced. Daróczy’s \( \beta \)-entropy for a partition \( \pi = \{P_1, \ldots, P_k\} \in \text{PART}(S) \) 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 \( \lim_{\beta \to 1} \mathcal{H}_\beta(\pi) = \text{Shannon’s entropy}\).

For \( \sigma, \pi \in \text{PART}(S) \), 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

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

we have

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

which yields the useful equivalent expression

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

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

\[
d_\beta(\pi, \sigma) = \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi)
\]

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

For \( \pi, \sigma \in \text{PART}(S) \) we write \( \pi \preceq \sigma \) 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 \( \pi \preceq \sigma \) and there is no partition \( \theta \in \text{PART}(S) \setminus \{\pi, \sigma\} \) such that \( \pi \preceq \theta \preceq \sigma \). This is equivalent to saying that \( \sigma \) is obtained from \( \pi \) by fusing together two blocks of \( \pi \). If \( \pi_1, \pi_2 \in \text{PART}(S) \), then we denote by \( \pi_1 \cdot \pi_2 \) 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 \( \text{PART}(S) \) is the partition \( \iota_S = \{ \{x\} \mid x \in S\} \) whose blocks are the singletons of \( S \); the largest partition of \( \text{PART}(S) \) 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 \( \mathcal{H}_B(\pi | \sigma) \geq \mathcal{H}_B(\pi | \sigma') \) and \( \sigma \leq \sigma' \) implies \( \mathcal{H}_B(\pi | \sigma) \leq \mathcal{H}_B(\pi | \sigma') \), 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 \( \text{adom}(B) \), 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 \( \pi = I_{\text{adom}(B)} \) observe that \( \pi_* = \kappa_B \).

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

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

Suppose that the list of objects sorted on the class 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 \( \text{adom}(B) \) as follows. A block of \( \pi_{B,A} \) consists of a maximal subsequence \( o_1[A], \ldots, o_k[A] \) of the previous sequence such that every object \( o_{k+1}, \ldots, o_n \) of this subsequence belongs to the same block \( K \) of the partition \( \kappa_A \). If \( x \in \text{adom}(B) \), 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:

<table>
<thead>
<tr>
<th>( o )</th>
<th>( B )</th>
<th>( A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( o_1 )</td>
<td>...</td>
<td>95.2</td>
</tr>
<tr>
<td>( o_2 )</td>
<td>...</td>
<td>110.1</td>
</tr>
<tr>
<td>( o_3 )</td>
<td>...</td>
<td>120.0</td>
</tr>
<tr>
<td>( o_4 )</td>
<td>...</td>
<td>125.5</td>
</tr>
<tr>
<td>( o_5 )</td>
<td>...</td>
<td>130.1</td>
</tr>
<tr>
<td>( o_6 )</td>
<td>...</td>
<td>140.0</td>
</tr>
<tr>
<td>( o_7 )</td>
<td>...</td>
<td>140.5</td>
</tr>
<tr>
<td>( o_8 )</td>
<td>...</td>
<td>168.2</td>
</tr>
<tr>
<td>( o_9 )</td>
<td>...</td>
<td>190.5</td>
</tr>
</tbody>
</table>

The partition \( \kappa_A \) has two blocks corresponding to the values ‘Y’ and ‘N’ and is given by \( \kappa_A = \{\{o_1, o_3, o_4, o_7, o_9\}, \{o_2, o_5, o_6\}\} \).

The partition \( \pi_{B,A} \) is:

\[ \pi_{B,A} = \{\{o_1\}, \{o_2, o_3, o_4, o_7, o_8, o_9\}\} \]

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 \( \mathcal{H}(\pi | \pi_B^T) \) the cutpoints \( t \) of \( T \) 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, naïve 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 \( \pi_B^T \) (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 Mántaras 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 \( T \) of cutpoints of \( \text{adom}(B) \) such that the blocks of \( \kappa_A \) are as pure as possible relative to the partition \( \pi_B^T \), that is, the conditional entropy \( \mathcal{H}_B(\kappa_A | \pi_B^T) \) 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 \( T \) is a set of cutpoints such that the conditional entropy \( \mathcal{H}_B(\kappa_A | \pi_B^T) \) is minimal among the set of cutpoints...
with the same number of elements, then $T$ consists of boundary points of the partition $\pi_{B,A}$ of $\text{dom}(B)$.

**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 $T$ is a set of cutpoints of $\text{dom}(B)$ such that the distance $d_\beta(\kappa_A, \pi_{B}^T)$ is minimal among the set of cutpoints with the same number of elements, then $T$ consists of boundary points of the partition $\pi_{B,A}$ of $\text{dom}(B)$.

**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 $\text{dom}(B)$ we shall seek a set of cutpoints $T$ such that $d_\beta(\kappa_A, \pi_{B}^T) = H_\beta(\kappa_A, \pi_{B}^T) + H_\beta(\pi_{B}^T | \kappa_A)$ is minimal. In other words, we shall seek a set of cutpoints such that the partition $\pi_{B}^T$ induced on the set of objects $S$ is as close as possible to the target partition $\kappa_A$.

Initially, before adding cutpoints, we have $T = \emptyset$, $\pi_{B}^T = \text{os} = \{S\}$, and therefore $H_\beta(\kappa_A, \text{os}) = H_\beta(\kappa_A)$. Observe that when the set $T$ grows the entropy $H_\beta(\kappa_A, \pi_{B}^T)$ decreases. Note that the use of conditional entropy $H_\beta(\kappa_A, \pi_{B}^T)$ tends to favor large cutpoint sets for which the partition $\pi_{B}^T$ is small in the partial ordered set $(\text{PART}(T), \leq)$. 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 $d_\beta(\kappa_A, \pi_{B}^T) = H_\beta(\kappa_A, \pi_{B}^T) + H_\beta(\pi_{B}^T | \kappa_A)$ the decrease of $H_\beta(\kappa_A, \pi_{B}^T)$ when the set of cutpoints grows is balanced by the increase in $H_\beta(\pi_{B}^T | \kappa_A)$. Note that initially we have $H_\beta(\text{os} | \kappa_A) = 0$. The discretization process can thus be halted when the distance $d_\beta(\kappa_A, \pi_{B}^T)$ stops decreasing. Thus, we retain as a set of cutpoints for discretization the set $T$ 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:**

1. compute the set $BP$ of boundary points of partition $\pi_{B,A}$; $T = \emptyset$; $d = \infty$;
2. while $BP \neq \emptyset$
   - let $t = \arg \min_{B \in BP} d_\beta(\kappa_A, \pi_{B}^{T \cup \{t\}})$;
   - if $d \geq d_\beta(\kappa_A, \pi_{B}^{T \cup \{t\}})$ then
     - begin
       - $T = T \cup \{t\}$;
       - $BP = BP - \{t\}$;
       - $d = d_\beta(\kappa_A, \pi_{B}^{T})$;
     - end
   - else
     - exit while loop;
   - end while
3. for $\pi_{B}^{T} = \{Q_0, \ldots, Q_\ell \}$ 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 $d_\beta(\kappa_A, \pi_{B}^{T \cup \{t\}})$ is less than the previous distance $d_\beta(\kappa_A, \pi_{B}^{T \cup \{t\}})$. 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 Fig. 1). Therefore, in practice we run a search for a new cutpoint only if $|d - d_\beta(\kappa_A, \pi_{B}^{T \cup \{t\}})| > 0.01d$.

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

- $\kappa_A = \{P_1, \ldots, P_n\}$,
- $\pi_{B}^{T} = \{Q_1, \ldots, Q_{m-2}, Q_{m-1} \cup Q_m\}$,
- $\pi_{B}^{T \cup \{t\}} = \{Q_1, \ldots, Q_{m-2}, Q_{m-1}, Q_m\}$.

![Fig. 1. Variation of distance with the size of the set of cutpoints.](image-url)
Since \( \beta > 1 \), by Equality (1), we have 
\[
d_p(\kappa_A, \pi_B) < d_p(\kappa_A, \pi_{B\prime})
\]
if and only if
\[
\sum_{i=1}^{n} |P_i| + \sum_{j=1}^{m} |Q_j| - 2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} |P_i \cap Q_j| < \sum_{i=1}^{n} |P_i| + \sum_{j=1}^{m-2} |Q_j| + |Q_{m-1} \cup Q_m| - 2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{m-2} |P_i \cap Q_j|,
\]
which is equivalent to
\[
|Q_{m-1}| + |Q_m| - 2 \cdot \sum_{i=1}^{n} |P_i \cap Q_{m-1}| - 2 \cdot \sum_{i=1}^{n} |P_i \cap Q_m| < |Q_{m-1} \cup Q_m| - 2 \cdot \sum_{i=1}^{n} (|P_i \cap Q_{m-1}| + |P_i \cap Q_m|),
\]
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 
\[
d_p(\kappa_A, \pi_B) \]
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 \( \pi_B \) 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.

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.”


### Table 2

<table>
<thead>
<tr>
<th>Database</th>
<th>Experimental results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discretization method</td>
</tr>
<tr>
<td>Heart-c</td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.5 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.8 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.9 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 2.0 )</td>
</tr>
<tr>
<td>Glass</td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.5 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.8 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.9 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 2.0 )</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.5 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.8 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.9 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 2.0 )</td>
</tr>
<tr>
<td>Iris</td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.5 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.8 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.9 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 2.0 )</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.8 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 1.9 )</td>
</tr>
<tr>
<td></td>
<td>( \beta = 2.0 )</td>
</tr>
</tbody>
</table>
fier 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 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.

Acknowledgment

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

Appendix A. Proofs of Theorems

A.1. Proof of Theorem 2.1

The proof is by induction on the number of cutpoints \( \ell = |T| \). If \( \ell = 0 \), the statement is immediate since in this case \( \pi^T_\beta \) is the one-class partition of the set of objects \( S \).

Suppose that the statement holds for set of cutpoints that contain \( \ell \) elements and let \( Z = T \cup \{ t \} \), where \( T = \{ t_1, \ldots, t_\ell \} \) is a set of cutpoints that is a subset of the set of boundary points of \( \pi^T_\beta \), \( |T| = \ell \) and \( t \not\in T \).

Let \( \kappa_A = \{ P_1, \ldots, P_k \} \) and \( \pi^T_\beta = \{ Q_0, \ldots, Q_\ell \} \), where \( \kappa_A \cap \pi^T_\beta \in \text{PART}(S) \). The conditional entropy \( H_\beta(\kappa_A|\pi^T_\beta) \) is given by:

![Table 2](image-url)  
Error rate for naïve Bayes classifiers

<table>
<thead>
<tr>
<th>Discretization method</th>
<th>Diabetes</th>
<th>Glass</th>
<th>Ionosphere</th>
<th>Iris</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta = 1.5 )</td>
<td>34.9</td>
<td>25.2</td>
<td>4.8</td>
<td>2.7</td>
</tr>
<tr>
<td>( \beta = 1.8 )</td>
<td>24.2</td>
<td>22.4</td>
<td>8.3</td>
<td>4</td>
</tr>
<tr>
<td>( \beta = 1.9 )</td>
<td>24.9</td>
<td>23.4</td>
<td>8.5</td>
<td>4</td>
</tr>
<tr>
<td>( \beta = 2.0 )</td>
<td>25.4</td>
<td>24.3</td>
<td>9.1</td>
<td>4.7</td>
</tr>
<tr>
<td>Weighted proportional</td>
<td>25.5</td>
<td>38.4</td>
<td>10.3</td>
<td>6.9</td>
</tr>
<tr>
<td>Proportional</td>
<td>26.3</td>
<td>33.6</td>
<td>10.4</td>
<td>7.5</td>
</tr>
</tbody>
</table>

![Table 3](image-url)  
Accuracy rate on test set on Khan’s data

<table>
<thead>
<tr>
<th>Discretization method</th>
<th>Accuracy rate on test set (%)</th>
<th>Misclassified “Noise” cases</th>
<th>Regular cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta = 1.3 )</td>
<td>76</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>( \beta = 1.35 )</td>
<td>60</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>( \beta = 1.4 )</td>
<td>84</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>( \beta = 1.5 )</td>
<td>80</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Since the partition $\mathbf{b}$ in order for $\mathbf{Qh}$ to achieve a local minimum. Thus, we have $\beta = 2$ is immediate since in this situation $F$ is a linear function of $\mu$. □

A.2 Proof of Theorem 2.2

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

The conditional entropy $\mathcal{H}(\pi_{B}^{\ell} | \mathbf{ka})$ is given by

$$
\mathcal{H}_{\mathbf{b}}(\pi_{B}^{\ell} | \mathbf{ka}) = \frac{1}{(1 - 2^{1-\beta})|S|^\beta}
\times \left( \sum_{i=1}^{\ell} |P_i| \beta - \sum_{j=0}^{k} \sum_{j=0}^{l} |P_i \cap Q_j| \beta \right).
$$

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

$$
\mathcal{H}_{\mathbf{b}}(\pi_{B}^{Z} | \mathbf{ka}) = \frac{1}{(1 - 2^{1-\beta})|S|^\beta}
\times \left( \sum_{i=1}^{\ell+1} |P_i| \beta - \sum_{j=0}^{k} \sum_{j=0}^{l} |P_i \cap Q_j| \beta \right).
$$

Thus, we have

$$
\mathcal{H}_{\mathbf{b}}(\pi_{B}^{Z} | \mathbf{ka}) - \mathcal{H}_{\mathbf{b}}(\pi_{B}^{\ell} | \mathbf{ka})
= \frac{1}{(1 - 2^{1-\beta})|S|^\beta}
\times \left( \sum_{i=1}^{\ell+1} |P_i| \beta - \sum_{j=0}^{k} \sum_{j=0}^{l} |P_i \cap Q_j| \beta \right).
$$

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

\[
\mathcal{H}_\beta(p^T_{\beta} | \kappa_4) - \mathcal{H}_\beta(p_{\beta} | \kappa_4) = \frac{1}{(1 - 2^{-1/\beta}) |S|^{\beta}} \times \left( H + |P_\beta \cap Q_\beta|^{1/\beta} - |P_\beta \cap Q'_\beta|^{1/\beta} - |P_\beta \cap Q''_\beta|^{1/\beta} \right),
\]

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

\[
\mathcal{H}_\beta(p^T_{\beta} | \kappa_4) - \mathcal{H}_\beta(p_{\beta} | \kappa_4) = \frac{1}{(1 - 2^{-1/\beta}) |S|^{\beta}} \left( H + n^{\beta} - \mu^{\beta} - (n - \mu)^{\beta} \right).
\]

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 \( d_\beta(\kappa_4, p^T_{\beta}) - d_\beta(\kappa_4, p_{\beta}) \) is the sum of the variations of the entropies \( \mathcal{H}_\beta(\kappa_4 | p^T_{\beta}) - \mathcal{H}_\beta(\kappa_4 | p_{\beta}) \) and \( \mathcal{H}_\beta(p^T_{\beta} | \kappa_4) - \mathcal{H}_\beta(p_{\beta} | \kappa_4) \). 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 \( t^l \) or with \( t^r \). □

References