DEPARTMENT OF COMPUTER SCIENCE
SERIES OF PUBLICATIONS C
REPORT C-2000-1

On the Splitting Properties of Common Attribute Evaluation Functions

Tapio Elomaa and Juho Rousu

UNIVERSITY OF HELSINKI
FINLAND

On the Splitting Properties of Common Attribute Evaluation Functions

Tapio Elomaa and Juho Rousu

Department of Computer Science
P. O. Box 26, FIN-00014 University of Helsinki, Finland

Series of Publications C, Report C-2000-1
Helsinki January 2000, 19 pages

Abstract

The complexity of numerical domain partitioning depends on the number of potential cut points. For a large family of attribute evaluation functions only boundary points need to be considered as candidates. We prove that an even more general property holds for many commonly-used functions. They do not obtain their optimal value within a segment of examples in which the relative class frequency distribution of examples is static. The borders of such segments are a subset of boundary points. Thus, even less cut points need to be examined for these functions.

The results shed a new light on the splitting properties of common attribute evaluation functions and they have practical value as well. The functions that are examined also include non-convex ones. Hence, the property introduced is not just another consequence of the convexity of a function.

Computing Reviews (1998) Categories and Subject Descriptors:
I.2.6 Learning [ARTIFICIAL INTELLIGENCE]: Concept learning, Induction, Knowledge acquisition

General Terms:
Learning, Analysis

Additional Key Words and Phrases:
numerical attributes, boundary points, decision trees.

1 Introduction

Fayyad and Irani (1992) showed that the *Average Class Entropy* and *Information Gain* functions (Quinlan 1986) obtain their optimal values for a numerical value range at a *boundary point*. Intuitively it means that these functions do not needlessly separate instances of the same class. The result reveals interesting fundamental properties of the functions, and it can also be put to use in practice: only boundary points need to be examined as potential cut points to recover the optimal binary split of the data.

Recently the utility of boundary points has been extended to cover other commonly-used evaluation functions and optimal multisplitting of numerical ranges (Elomaa and Rousu 1999). Other recent studies concerning the splitting properties of attribute evaluation functions include Breiman's (1996) research of the characteristics of ideal partitions of some impurity functions and Codrington and Brodley's (2000) study of the general requirements of well-behaved splitting functions.

This paper continues to explore the splitting properties of attribute evaluation functions. We introduce a generalized version of boundary points—the so-called *segment borders*—which exclude all cut points in the numerical range that separate subsets of identical relative class frequency distributions. The separated subsets do not need to be class uniform to warrant the exclusion, as is the case with boundary points.

We show that it suffices to examine segment borders in optimizing the value of the best-known attribute evaluation functions. Hence, the changes in class distribution, rather than relative impurities of the subsets, define the potential locations of the optimal cut points (cf. López de Màntaras 1991). Two of the examined functions are non-convex. Hence, the property of splitting on segment borders is not only a consequence of the convexity of a function.

A *partition* $\biguplus_{i=1}^{k} S_i$ of the sample S consists of k non-empty, disjoint subsets and covers the whole domain. When splitting a set S of examples on the basis of the value of an attribute A, there is a set of thresholds $\{T_1, \ldots, T_{k-1}\} \subseteq \mathrm{Dom}(A)$ that defines a partition $\biguplus_{i=1}^{k} S_i$ for the sample in an obvious manner:

$$S_i = \begin{cases} \{s \in S \mid \mathrm{val}_A(s) \le T_1\} & \text{if } i = 1, \\ \{s \in S \mid T_{i-1} < \mathrm{val}_A(s) \le T_i\} & \text{if } 1 < i < k, \\ \{s \in S \mid \mathrm{val}_A(s) > T_{k-1}\} & \text{if } i = k, \end{cases}$$

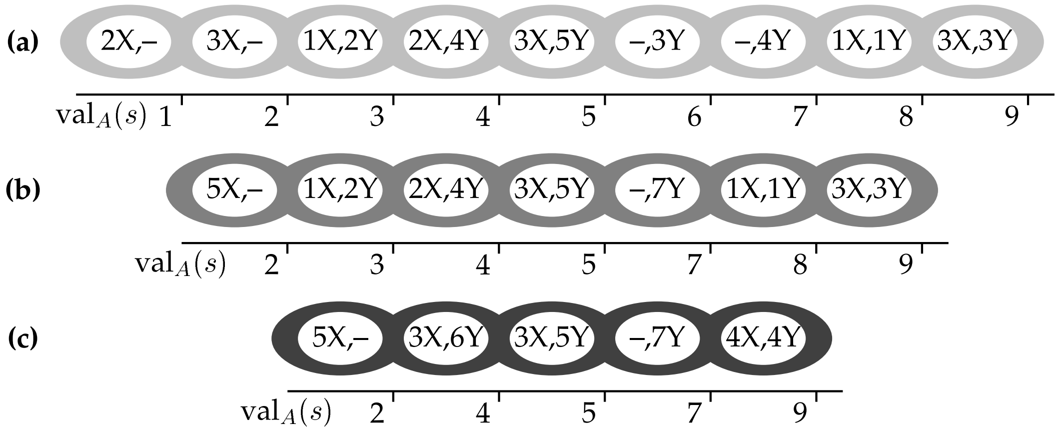

Figure 1: The (a) bins, (b) blocks, and (c) segments in the domain of a numerical attribute A in a hypothetical sample.

where $\mathrm{val}_A(s)$ denotes the value of attribute A in example s.

We take all logarithms in this paper to be natural logarithms; it makes the manipulation and notation simpler. It is easy to check that our proofs can be worked through with binary logarithms as well. The classification of an example s is its value for the class attribute C, $\mathrm{val}_C(s)$.

This paper is organized as follows. Section 2 recapitulates boundary points and introduces example segments. In Section 3 we prove that six well-known functions do not partition within a segment. Empirical evaluation of the utility of segment borders with 28 UCI test domains is presented in Section 4. Finally, Section 5 relates the results to those of Breiman (1996) and outlines further research directions.

2 Example segments

We recapitulate bins and blocks of examples as well as boundary points. Furthermore, we introduce segments. Rather than give unnecessarily complicated formal definitions for these simple concepts, we present them intuitively with the help of an illustration.

The definition of boundary points assumes that the sample has been sorted into ascending order with respect to the value of the numerical attribute under consideration (Fayyad and Irani 1992). Sorting is a typical preprocessing step in attribute evaluation strategies. It produces a sequence of *bins*, where all examples with an equal value for the attribute in question make up a bin of examples. Bin borders are the possible cut points of the value range. In Fig. 1a a hypothetical sample has been arranged into bins with respect to an integer-valued attribute A. The numbers of instances of different classes (X and Y in this case) belonging to the bins are depicted by the figures within them. To determine the correlation between the value of an attribute and that of the class it suffices to examine their mutual frequencies.

To construct *blocks* of examples we merge adjacent class uniform bins with the same class label (see Fig. 1b). The boundary points of the example sequence are the borders of its blocks. The points obtained thus are exactly the same as those that come out from the definition of Fayyad and Irani (1992). Block construction leaves all bins with a mixed class distribution as their own blocks.

From bins we obtain *segments* of examples by combining adjacent bins with an equal relative class distribution (see Fig. 1c). Segments group together adjacent mixed-distribution bins that have equal relative class distribution. Also adjacent class uniform bins fulfill this condition; hence, uniform blocks are a special case of segments and segment borders are a subset of boundary points.

The rest of this paper is concerned with showing that many attribute evaluation functions obtain their optimal value on segment borders.

3 Most common evaluation functions split on segment borders

All the following proofs have the same setting. The sample S contains three subsets, P, Q, and R with class frequency distributions

$$p = \sum_{j=1}^{m} p_j,\ q = \sum_{j=1}^{m} q_j,\ \text{and } r = \sum_{j=1}^{m} r_j,$$

where p is the number of examples in P and p_j is the number of instances of class j in P. Furthermore, m is the number of classes. The notation is similar also for Q and R. Let us define $w_j = q_j/q \in [0, 1]$.

We consider the k-ary partition $\biguplus_{i=1}^{k} S_i$ of the sample S, where subsets S_h and S_{h+1} consist of the set $P \cup Q \cup R$, so that the split point is inside Q,

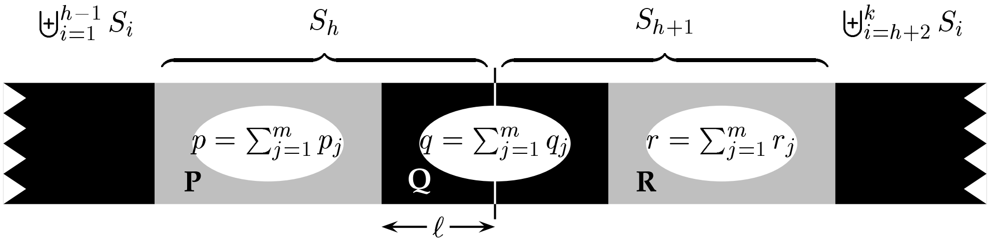

Figure 2: The following proofs consider partitioning of the example set $P \cup Q \cup R$ into two subsets S_h and S_{h+1} within Q. No matter where, within Q, the cut point is placed, equal class distributions result.

on the border of P and Q, or that of Q and R (see Fig. 2). Let ℓ be an integer, $0 \le \ell \le q$. We assume that splitting the set Q so that ℓ examples belong to S_h and $q - \ell$ to S_{h+1} results in identical class frequency distributions for both subsets of Q regardless of the value of ℓ. In other words, for all j and ℓ it holds that $q_j(\ell) = w_j \cdot \ell$ where $q_j(\ell)$ is the frequency of class j in S_h.

The proofs treat the evaluation functions and their component functions as continuous in $[0, q]$ and twice differentiable, even though they are defined to be discrete. Observe that this causes no harm, since we only consider proving the *absence* of certain local extremas.

The proofs show in the multisplitting situation that the cut point in between two arbitrarily chosen partition subsets S_h and S_{h+1} is on a segment border. The remaining partition subsets are not affected by the placement of the cut point within $S_h \cup S_{h+1}$. Therefore, their impact usually disappears when the proof involves differentiation of the function.

The remainder of this section shows explicitly in our proof setting that six commonly-used attribute evaluation functions do not obtain their optimal values within a segment.

3.1 Average Class Entropy

The Average Class Entropy of the partition $\biguplus_i S_i$ is

$$ACE\left(\biguplus_i S_i\right) = \sum_i \frac{|S_i|}{|S|} H(S_i) = \frac{1}{|S|} \sum_i |S_i| H(S_i),$$

where H is the *entropy* function,

$$H(S) = -\sum_{j=1}^{m} P(C_j, S) \log P(C_j, S),$$

in which m denotes the number of classes and $P(C, S)$ stands for the proportion of the examples in S that belong to the class C.

Theorem 1 *The Average Class Entropy optimal partitions are defined on segment borders.*

Proof Let $L(\ell)$ denote the value of $\sum_{i=1}^{h} |S_i| H(S_i)$ when S_h contains P and the first ℓ examples from Q, and $R(\ell)$ the value $\sum_{i=h+1}^{k} |S_i| H(S_i)$. Now,

$$\begin{aligned} L(\ell) &= \sum_{i=1}^{h-1} |S_i| H(S_i) - \sum_{j=1}^{m} (p_j + w_j \ell) \log \frac{p_j + w_j \ell}{p + \ell} \\ &= \sum_{i=1}^{h-1} |S_i| H(S_i) + (p + \ell) \log(p + \ell) - \sum_{j=1}^{m} (p_j + w_j \ell) \log(p_j + w_j \ell) \end{aligned}$$

and, similarly,

$$\begin{aligned} R(\ell) &= (r + q - \ell) \log(r + q - \ell) - \sum_{j=1}^{m} (r_j + q_j - w_j \ell) \log(r_j + q_j - w_j \ell) \\ &\quad + \sum_{i=h+2}^{k} |S_i| H(S_i). \end{aligned}$$

Since the first sum in the formula of $L(\ell)$ is independent of the placing of the h-th cut point, it differentiates to zero and the second derivative of $L(\ell)$ is

$$\begin{aligned} L''(\ell) &= \frac{\mathrm{d}}{\mathrm{d}\ell} \left(\log(p + \ell) - \sum_{j=1}^{m} w_j \log(p_j + w_j \ell) \right) \\ &= \frac{1}{p + \ell} - \sum_{j=1}^{m} \frac{w_j^2}{p_j + w_j \ell}. \end{aligned}$$

We apply $w_j/(p_j + w_j \ell) = 1/(p_j/w_j + \ell)$ and obtain

$$L''(\ell) = \frac{1}{p + \ell} - \sum_{j=1}^{m} w_j \frac{1}{p_j/w_j + \ell}.$$

The sum can be interpreted as the weighted arithmetic mean of the terms $1/(p_j/w_j+\ell), 1 \le j \le m$, and by the arithmetic-harmonic mean inequality (Hardy, Littlewood, and Pólya, 1934; Meyer, 1984) be bound from below by the corresponding harmonic mean

$$\begin{aligned}\sum_{j=1}^{m} w_j \frac{1}{p_j/w_j+\ell} &\ge \frac{1}{\sum_{j=1}^{m} w_j (p_j/w_j+\ell)} \\ &= \frac{1}{\sum_{j=1}^{m}(p_j+w_j\ell)} \\ &= \frac{1}{p+\ell}.\end{aligned}$$

Thus, we obtain

$$L''(\ell) \le \frac{1}{p+\ell} - \frac{1}{p+\ell} = 0.$$

Correspondingly, the second derivative of $R(\ell)$ can be approximated by majorizing the second term by the harmonic mean

$$\begin{aligned}R''(\ell) &= \frac{1}{r+q-\ell} - \sum_{j=1}^{m} \frac{w_j^2}{r_j+q_j-w_j\ell} \\ &= \frac{1}{r+q-\ell} - \sum_{j=1}^{m} w_j \frac{1}{(r_j+q_j)/w_j-\ell} \\ &\le \frac{1}{r+q-\ell} - \frac{1}{\sum_{j=1}^{m} w_j((r_j+q_j)/w_j-\ell)} \\ &= \frac{1}{r+q-\ell} - \frac{1}{\sum_{j=1}^{m}(r_j+q_j-w_j\ell)} \\ &= 0.\end{aligned}$$

Hence, we have shown that the second derivative of *ACE*, $ACE''(\ell) = (L''(\ell)+R''(\ell))/|S|$ is non-positive for all ℓ. This forces all local extrema of *ACE* within Q to be maxima. ■

3.2 Information Gain

Information gain function, or the *Mutual Information*, is a simple modification of *ACE*. Thus, proving that it does not partition within segments is straightforward.

Theorem 2 *The Information Gain optimal partitions are defined on segment borders.*

Proof The Information Gain of the partition $\biguplus_{i=1}^{k} S_i$, when the h-th cut point is placed after the ℓ-th example of Q, is

$$IG(\ell) = H(S) - ACE(\ell).$$

The constant term $H(S)$ that does not depend on the value of ℓ differentiates to zero. Therefore, $IG'(\ell) = -ACE'(\ell)$ and its second derivative is $-ACE''(\ell)$. From the proof of Theorem 1 we know that $ACE''(\ell) \leq 0$, which means that $IG''(\ell) \geq 0$. Hence, IG cannot have a local maximum within segment Q. ■

3.3 Gain Ratio

To penalize against *IG*'s excessive favoring of multi-valued nominal attributes and multisplitting numerical attribute value ranges, Quinlan (1986) suggested dividing the *IG* score of a partition by the term

$$\kappa\left(\biguplus_i S_i\right) = -\sum_i \frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|}.$$

The resulting evaluation function is the *Gain Ratio*

$$GR\left(\biguplus S_i\right) = IG\left(\biguplus S_i\right) / \kappa\left(\biguplus S_i\right).$$

The *IG* function was already inspected above. Therefore, the following proof concentrates on the denominator κ.

Theorem 3 *The Gain Ratio optimal partitions are defined on segment borders.*

Proof The denominator κ of the *GR* formula in our proof setting is

$$\begin{aligned}
\kappa(\ell) &= \kappa\left(\biguplus_{i=1}^{h-1} S_i\right) - \frac{1}{|S|}\Big(|S_h|(\log|S_h| - \log|S|) \\
&\quad + |S_{h+1}|(\log|S_{h+1}| - \log|S|)\Big) + \kappa\left(\biguplus_{i=h+2}^{k} S_i\right) \\
&= \kappa\left(\biguplus_{i=1}^{h-1} S_i\right) + \frac{1}{|S|}\Big((p+q+r)\log|S| - (p+\ell)\log(p+\ell) \\
&\quad - (r+q-\ell)\log(r+q-\ell)\Big) + \kappa\left(\biguplus_{i=h+2}^{k} S_i\right).
\end{aligned}$$

The first and the last term are independent of ℓ and differentiate to zero. Thus, the second derivative of $\kappa(\ell)$ with respect to ℓ is

$$\begin{aligned}\kappa''(\ell) &= \frac{\mathrm{d}}{\mathrm{d}\ell}\left(\frac{1}{|S|}(-\log(p+\ell)-\log(r+q-\ell))\right)\\ &= \frac{1}{|S|}\left(\frac{-1}{p+\ell}+\frac{-1}{r+q-\ell}\right)\\ &< 0. \end{aligned}\tag{1}$$

The first derivative of $GR(\ell)$ is given by

$$GR'(\ell)=\frac{IG'(\ell)\kappa(\ell)-\kappa'(\ell)IG(\ell)}{\kappa^2(\ell)}.$$

Let us define $N(\ell)=IG'(\ell)\kappa(\ell)-\kappa'(\ell)IG(\ell)$, and note that

$$\begin{aligned}N'(\ell) &= IG''(\ell)\kappa(\ell)+\kappa'(\ell)IG'(\ell)-\kappa''(\ell)IG(\ell)-\kappa'(\ell)IG'(\ell)\\ &= IG''(\ell)\kappa(\ell)-\kappa''(\ell)IG(\ell)\\ &\geq 0,\end{aligned}$$

because for each $0<\ell<q$ it holds by definition that $\kappa(\ell)>0$ and $IG(\ell)\geq 0$. Furthermore, by Theorem 2 we know that $IG''(\ell)\geq 0$ and by Eq. 1 that $\kappa''(\ell)<0$.

Now the second derivative of $GR(\ell)$ is expressed by

$$GR''(\ell)=\frac{N'(\ell)\kappa^2(\ell)-2\kappa(\ell)\kappa'(\ell)N(\ell)}{\kappa^4(\ell)}.$$

Let $\psi\in]0,q[$ be a potential location for a local maximum of GR, i.e., such a point that $GR'(\psi)=0$. Then also $N(\psi)=0$ and the expression for $GR''(\psi)$ is further simplified to

$$GR''(\psi)=N'(\psi)/\kappa^2(\psi),$$

which is larger than zero because $N'(\psi)\geq 0$ and $\kappa^2(\psi)>0$. In other words, $GR(\psi)$ is not a local maximum. Since ψ was chosen arbitrarily, we have shown that $GR(\ell)$ can only obtain its maximum value when the threshold is placed at either of the segment borders, where $\ell=0$ and $\ell=q$, respectively. ■

3.4 Normalized Distance Measure

The *Normalized Distance Measure* was proposed by López de Màntaras (1991) as an alternative to the Information Gain and Gain Ratio functions. It can be expressed with the help of the Information Gain as

$$\mathit{ND}\left(\biguplus S_i\right) = 1 - \mathit{IG}\left(\biguplus S_i\right) / \lambda\left(\biguplus S_i\right),$$

where

$$\lambda\left(\biguplus_{i=1}^{k} S_i\right) = -\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{M(j,S_i)}{|S|} \log \frac{M(j,S_i)}{|S|},$$

in which $M(j,S)$ stands for the number of instances of class j in the set S.

The following proof concerns instead the function

$$\mathit{ND}_1\left(\biguplus S_i\right) = 1 - \mathit{ND}\left(\biguplus S_i\right) = \mathit{IG}\left(\biguplus S_i\right) / \lambda\left(\biguplus S_i\right),$$

from which the claim directly follows for *ND*.

The ND_1 formula resembles that of *GR*. Therefore, the proof outline is also the same.

Theorem 4 *The Normalized Distance Measure optimal partitions are defined on segment borders.*

Proof Let $L(\ell)$ denote the value of $\lambda(\biguplus_{i=1}^{h} S_i)$ and $R(\ell)$ the value $\lambda(\biguplus_{i=h+1}^{k} S_i)$.

$$\begin{aligned} L(\ell) &= \lambda\left(\biguplus_{i=1}^{h-1} S_i\right) - \sum_{j=1}^{m} \frac{p_j + w_j\ell}{|S|} \log \frac{p_j + w_j\ell}{|S|} \\ &= \lambda\left(\biguplus_{i=1}^{h-1} S_i\right) + \frac{1}{|S|}\left((p+\ell)\log|S| - \sum_{j=1}^{m}(p_j + w_j\ell)\log(p_j + w_j\ell)\right) \end{aligned}$$

and

$$\begin{aligned} R(\ell) &= \frac{1}{|S|}\left((r+q-\ell)\log|S| - \sum_{j=1}^{m}(r_j + q_j + w_j\ell)\log(r_j + q_j - \ell)\right) \\ &\quad + \lambda\left(\biguplus_{i=h+2}^{k} S_i\right). \end{aligned}$$

The second derivative of $L(\ell)$ is given by

$$\begin{aligned} L''(\ell) &= \frac{\mathrm{d}}{\mathrm{d}\ell}\frac{1}{|S|}\left(\log|S| - \sum_{j=1}^{m} w_j \log(p_j + w_j\ell)\right) \\ &= \frac{-1}{|S|}\sum_{j=1}^{m}\frac{w_j^2}{p_j + w_j\ell} \\ &\leq 0, \end{aligned}$$

because $|S|$, w_j, p_j, and ℓ are all non-negative.

Correspondingly, the second derivative of $R(\ell)$ is

$$R''(\ell) = \frac{-1}{|S|}\sum_{j=1}^{m}\frac{w_j^2}{r_j + q_j - w_j\ell} \leq 0.$$

Thus, we have proved that the second derivative of λ, $\lambda''(\ell) = L''(\ell) + R''(\ell)$ is non-positive for all ℓ.

The first derivative of $ND_1(\ell)$ is given by

$$ND_1'(\ell) = \frac{IG'(\ell)\lambda(\ell) - \lambda'(\ell)IG(\ell)}{\lambda^2(\ell)}.$$

Let us define $K(\ell) = IG'(\ell)\lambda(\ell) - \lambda'(\ell)IG(\ell)$, and note that

$$\begin{aligned} K'(\ell) &= IG''(\ell)\lambda(\ell) + \lambda'(\ell)IG'(\ell) - \lambda''(\ell)IG(\ell) - \lambda'(\ell)IG'(\ell) \\ &= IG''(\ell)\lambda(\ell) - \lambda''(\ell)IG(\ell) \\ &\geq 0, \end{aligned}$$

because for each $0 < \ell < q$ it holds by definition that $\lambda(\ell) > 0$ and $IG(\ell) \geq 0$. Furthermore, by Theorem 2 we know that $IG''(\ell) \geq 0$ and by the above inspection that $\lambda''(\ell) \leq 0$.

Now the second derivative of $ND_1(\ell)$ is expressed by

$$ND_1''(\ell) = \frac{K'(\ell)\lambda^2(\ell) - 2\lambda(\ell)\lambda'(\ell)K(\ell)}{\lambda^4(\ell)}.$$

Let $\psi \in]0, q[$ be a potential location for a local maximum of ND_1, i.e., such a point that $ND_1'(\psi) = 0$. Then also $K(\psi) = 0$ and the expression for $ND_1''(\psi)$ is further simplified to

$$ND_1''(\psi) = K'(\psi)/\lambda^2(\psi),$$

which is larger than zero because $K'(\psi) \geq 0$ and $\lambda^2(\psi) > 0$. In other words, $ND_1(\psi)$ is not a local maximum. Since ψ was chosen arbitrarily, we have shown that $ND_1(\ell)$ can only obtain its maximum value when the threshold is placed at either of the segment borders, where $\ell = 0$ and $\ell = q$, respectively.

Thus, the local extrema of ND_1 within Q are minima, which makes them local maxima of $ND(\ell) = 1 - ND_1(\ell)$. Hence, Normalized Distance measure does not obtain its minimum value within a segment. ■

3.5 Gini Index

Gini Index (of diversity), or the *Quadratic Entropy*, (Breiman et al., 1984; Breiman, 1996) is defined as

$$GI\left(\biguplus_i S_i\right) = \sum_i \frac{|S_i|}{|S|} gini(S_i),$$

in which *gini* is the impurity measure

$$gini(S) = \sum_{j=1}^{m} P(C_j, S)(1 - P(C_j, S)) = 1 - \sum_{j=1}^{m} P^2(C_j, S),$$

where $P(C, S)$ denotes the proportion of instances of class C in the data S.

Theorem 5 *The Gini Index optimal partitions are defined on segment borders.*

Proof Let $L(\ell)$ denote the value of $\sum_{i=1}^{h} |S_i| gini(S_i)$ when S_h contains P and the first ℓ examples from Q, and $R(\ell)$ the value $\sum_{i=h+1}^{k} |S_i| gini(S_i)$. Now,

$$\begin{aligned} L(\ell) &= \sum_{i=1}^{h-1} |S_i| gini(S_i) + \sum_{j=1}^{m} (p_j + w_j \ell)\left(1 - \frac{p_j + w_j \ell}{p + \ell}\right) \\ &= \sum_{i=1}^{h-1} |S_i| gini(S_i) + (p + \ell) - \sum_{j=1}^{m} \frac{(p_j + w_j \ell)^2}{p + \ell}. \end{aligned}$$

The second derivative of $L(\ell)$ is

$$L''(\ell) = \frac{\mathrm{d}}{\mathrm{d}\ell}\left(1 - \sum_{j=1}^{m} \frac{2(p_j + w_j \ell) w_j (p + \ell) - (p_j + w_j \ell)^2}{(p + \ell)^2}\right)$$

$$
\begin{aligned}
&= \frac{\mathrm{d}}{\mathrm{d}\ell}\left(1 - \sum_{j=1}^{m}\frac{2(p_j + w_j\ell)w_j}{(p+\ell)} + \sum_{j=1}^{m}\frac{(p_j + w_j\ell)^2}{(p+\ell)^2}\right) \\
&= -\sum_{j=1}^{m}\frac{2w_j^2(p+\ell) - 2(p_j + w_j\ell)w_j}{(p+\ell)^2} \\
&\quad + \sum_{j=1}^{m}\frac{2(p_j + w_j\ell)w_j(p+\ell)^2 - 2(p+\ell)(p_j + w_j\ell)^2}{(p+\ell)^4} \\
&= \frac{1}{(p+\ell)^3}\sum_{j=1}^{m}\Big(2(p_j + w_j\ell)w_j(p+\ell) - 2w_j^2(p+\ell)^2 \\
&\quad + 2(p_j + w_j\ell)w_j(p+\ell) - 2(p_j + w_j\ell)^2\Big) \\
&= (p+\ell)^{-3}\sum_{j=1}^{m}\Big(4(p_j + w_j\ell)w_j(p+\ell) - 2w_j^2(p+\ell)^2 - 2(p_j + w_j\ell)^2\Big) \\
&= (p+\ell)^{-3}\sum_{j=1}^{m}\Big(4p_jw_jp + 4p_jw_j\ell + 4w_j^2\ell p + 4w_j^2\ell^2 \\
&\quad - 2w_j^2p^2 - 4w_j^2p\ell - 2w_j^2\ell^2 - 2p_j^2 - 4p_jw_j\ell - 2w_j^2\ell^2\Big) \\
&= (p+\ell)^{-3}\sum_{j=1}^{m}\Big(4p_jw_jp - 2w_j^2p^2 - 2p_j^2\Big) \\
&= -2(p+\ell)^{-3}\sum_{j=1}^{m}(p_j + w_jp)^2 \\
&\leq 0.
\end{aligned}
$$

By symmetry we determine that $R''(\ell) \leq 0$ as well. Thus, $GI''(\ell) = (L''(\ell) + R''(\ell))/|S| \leq 0$ and, therefore, *GI* does not obtain its minimum value within the segment Q. ■

3.6 Training Set Error

The *majority* class of sample S is its most frequently occurring class:

$$\mathrm{maj}_C(S) = \arg\max_{1\leq j\leq m} |\{s \in S \mid \mathrm{val}_C(s) = j\}|.$$

The number of *disagreeing* instances, those in the set S not belonging to its majority class, is given by

$$\delta(S) = |\{s \in S \mid \mathrm{val}_C(s) \neq \mathrm{maj}_C(S)\}|.$$

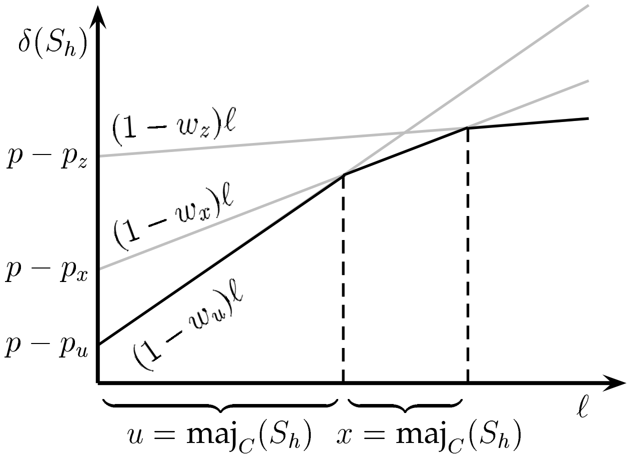

Figure 3: The number of instances of other classes than j grows linearly with increasing ℓ for all j. $\delta(S_h)$ is convex.

Training Set Error is the number of training instances falsely classified in the partition. For a partition $\biguplus S_i$ of S it is defined as

$$TSE\left(\biguplus_i S_i\right) = \sum_i \delta(S_i).$$

The number of instances in S_h from other classes than j, $(p - p_j) + (1 - w_j)\ell$, is linearly increasing for any j, since the first term is constant and $0 \leq w_j \leq 1$. Respectively, in S_{h+1} the number of those instances, $(r - r_j) + (1 - w_j)(q - \ell)$, decreases with increasing ℓ.

In our proof setting, the majority class of S_h depends on the growth rates of classes in Q and the number of their instances in P. First, when $\ell = 0$, the majority class of P, say u, is also the majority class of S_h. Subsequently an other class x, with strictly larger growth rate $w_x > w_u$ may become the majority class of S_h (see Fig. 3). Observe that $p_x \leq p_u$. As a combination of non-decreasing functions, $\delta(S_h)$ is also non-decreasing.

Theorem 6 *The Training Set Error optimal partitions are defined on segment borders.*

Proof Let us examine the value of *TSE*$(l) = \delta(S_h) + \delta(S_{h+1})$ at an arbitrary cut point $\ell = l$, $0 \leq l \leq q$. Let u and v be the majority classes of S_h and

S_{h+1}, respectively, in this situation. Then,

$$TSE(l) = (p - p_u) + (1 - w_u)l + (r - r_v) + (1 - w_v)(q - l).$$

We now show that a smaller training set error is obtained by moving the cut point to the left or to the right from l. There are four possible scenarios for the changes of majority classes of S_h and S_{h+1} when the cut point is moved: (i) neither of them changes, only the majority class of (ii) S_h or (iii) S_{h+1} changes, or (iv) both of them change. Let x and y, when needed, be the new majority classes of S_h and S_{h+1}, respectively.

Assume, now, that $w_u \leq w_v$. Let us consider the four scenarios mentioned when moving the cut point one example to the left.

(i)
$$\begin{aligned} TSE(l-1) &= (p - p_u) + (1 - w_u)(l - 1) + (r - r_v) + (1 - w_v)(q - l + 1) \\ &= TSE(l) + w_u - w_v \\ &\leq TSE(l), \end{aligned}$$

because $w_u - w_v \leq 0$ by the assumption.

(ii) The majority class of S_h becomes x and v remains to be the majority class of S_{h+1}. Then

$$\begin{aligned} TSE(l-1) &= (p - p_x) + (1 - w_x)(l - 1) \\ &\quad + (r - r_v) + (1 - w_v)(q - l + 1) \\ &\leq TSE(l) + w_x - w_v, \end{aligned}$$

because $(p-p_x) \leq (p-p_u)$ and $(1-w_x) < (1-w_u)$. Since $w_x < w_u \leq w_v$ by the assumption, we have shown that $TSE(l-1) \leq TSE(l)$.

(iii) The majority class of S_h remains to be u and y becomes the majority class of S_{h+1}. Observe that then $(r - r_y) + (1 - w_y)(q - l + 1) \leq (r - r_v) + (1 - w_v)(q - l + 1)$, by y being the majority class of S_{h+1}. Thus,

$$TSE(l-1) \leq TSE(l) + w_u - w_v \leq TSE(l).$$

(iv) If both majority classes change, then by combining (ii) and (iii) we see that $TSE(l-1) \leq TSE(l)$.

Hence, in all scenarios a smaller value of *TSE* is obtained by moving the cut point.

Respectively, if $w_u \geq w_v$ we can obtain a smaller training set error for $S_h \uplus S_{h+1}$ by sliding the cut point forward in Q.

$$\begin{aligned}\text{(i) } TSE(l+1) &= (p-p_u)+(1-w_u)(l+1)+(r-r_v)+(1-w_v)(q-l-1)\\ &= TSE(l)+w_v-w_u\\ &\leq TSE(l),\end{aligned}$$

because $w_v - w_u \leq 0$ by the assumption.

(ii) The majority class of S_h becomes x and v remains to be the majority class of S_{h+1}. Observe that then $(p-p_x)+(1-w_x)(l+1) \leq (p-p_u)+(1-w_u)(l+1)$, by x being the majority class of S_h. Thus,

$$TSE(l+1) \leq TSE(l)+w_v-w_u \leq TSE(l).$$

(iii) The majority class of S_h remains to be u and y becomes the majority class of S_{h+1}. Then

$$\begin{aligned}TSE(l+1) &= (p-p_u)+(1-w_u)(l+1)\\ &\quad +(r-r_y)+(1-w_y)(q-l-1)\\ &\leq TSE(l)+w_y-w_u,\end{aligned}$$

because $(r-r_y) \leq (r-r_v)$ and $(1-w_y) < (1-w_v)$. Since $w_y < w_v \leq w_u$ by the assumption, we have shown that $TSE(l+1) \leq TSE(l)$.

(iv) If both majority classes change, then by combining (ii) and (iii) we see that $TSE(l+1) \leq TSE(l)$.

In any case, the cut point can be slid all the way to one of the borders of Q. Because l was chosen arbitrarily and $TSE(\biguplus_{i=1}^{k} S_i) = \sum_{i=1}^{h-1} \delta(S_i) + TSE(l) + \sum_{i=h+2}^{k} \delta(S_i)$, we have proved the claim. ■

4 Empirical evaluation

We test for 28 well-known UCI domains (Blake and Merz, 1998) the effects of concentrating on segment borders. Fig. 4 depicts the average numbers

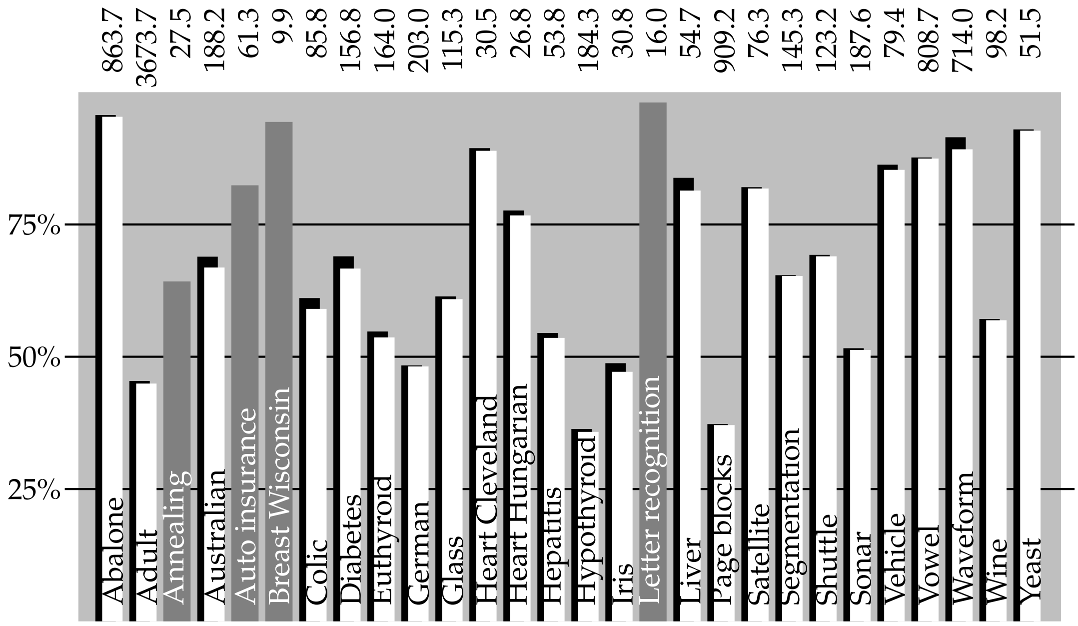

Figure 4: The average number of bin borders (the figures on top) and the relative numbers of boundary points (black bars) and segment borders (white bars) per numerical attribute of the domain. Gray bars indicate that the numbers of segment borders and boundary points are the same.

of bin borders (the figures on the right) and the relative portions of boundary points (black bars) and segment borders (white bars) per numerical attribute of the domain. Gray bars indicate that the numbers of boundary points and segment borders are the same.

From Fig. 4 we see that the average number of segment borders per attribute is only marginally smaller than that of the boundary points. By combining bins into segments, in real-world data, almost all reduction in the number of points that need to be examined comes from combination of class uniform bins, only very few mixed bins get combined.

The segment construction is as efficient as block combination. Therefore, nothing is lost by taking advantage of the small reduction in the number of cut points examined.

5 Discussion

We have shown that the result of Fayyad and Irani (1992) can be properly generalized. Class uniform bins are not the only ones that can be grouped together without losing the optimal partition. In practice, though, they turn out to be far more numerous than other segments with static relative class frequency distribution. However, even small reductions in the number of cut points are valuable in the optimal partitioning tasks, where the time complexity can be quadratic or exponential in the number of cut points.

Most common evaluation functions behave in the way we would like them to: minimizing on non-boundary cut points would mean needless separation of instances of the same class. Moreover, minimizing within segments that have an identical relative class distributions would mean separating instances even if we have no evidence that they need different handling.

Even if the popular evaluation functions are similar in the above sense, their biases are different. Breiman (1996) studied for Gini Index and Average Class Entropy which class frequency distribution, if given the freedom to choose any, would produce the optimal score for a binary partition. He showed that Gini Index ideally separates all the instances of the majority class from the rest of the examples. Entropy minimization, on the other hand, aims at producing equal-sized subsets. In practice, the choices for the class frequency distributions are limited by the sample characteristics and the ideal partitions cannot necessarily be realized.

In Breiman's (1996) setting α_j denotes the proportion of examples of class j that lie to the left of a binary split. Hence, the vector $\alpha = \langle \alpha_1, \ldots, \alpha_m \rangle$ is a point in the hypercube $A = [0, 1]^m$. Moving the cut point within an example segment corresponds to moving vector α on a straight line in the hypercube A. Moving it over a set that spans multiple segments, forms a piecewise linear trajectory in A. The segment borders are the turning points of the trajectory (see Fig. 5). If the example segment is class uniform, the line is axis-parallel. Non-boundary cut points fall on such line segments. Boundary points other than segment borders are situated on lines that are not axis-parallel, which correspond to segments with a mixed class distribution.

The practical uses of the results in this paper are somewhat hampered by the fact that small differences in neighboring blocks are inevitably present

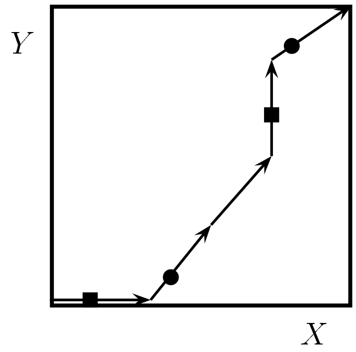

Figure 5: The effects on vector α of moving the cut point through the data set in Fig. 1 in Breiman's model. The discs denote boundary points other than segment borders and the squares denote non-boundary cut points.

even if their underlying true class distributions were the same. These differences arise because of sampling of examples and noise. Hence, it is rare to find a sequence of cut points to lie exactly on a single line in the space A.

Thus, it would be useful to consider situations where the the relative class distributions of the neighboring blocks were allowed to differ. The questions for further research include whether the absence of optima can still be guaranteed, how much deviation can be allowed, and which types of deviations make it easier to guarantee the absence of optima within the example segment.

References

Blake, C. L. and Merz, C. J. (1998): UCI repository of machine learning databases. Irvine, CA: Univ. of California, Dept. of Information and Computer Science. `http://www.ics.uci.edu/~mlearn/MLRepository.html`

Breiman, L. (1996). Some properties of splitting criteria. *Machine Learning* **24**: 41–47.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

Codrington, C. and Brodley, C. (forthcoming). On the qualitative behavior of impurity-based splitting rules I: The minima-free property. *Machine Learning*.

Elomaa, T. and Rousu, J. (1999). General and efficient multisplitting of numerical attributes. *Machine Learning* **36**: 201–244.

Fayyad, U. and Irani, K. (1992). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning* **8**: 87–102.

Hardy, G. H., Littlewood, J. E., and Pólya, G. (1934). *Inequalities*. Cambridge, UK: Cambridge Univ. Press.

López de Màntaras, R. (1991). A distance-based attribute selection measure for decision tree induction. *Machine Learning* **6**: 81–92.

Meyer, B. (1984). Some inequalities for elementary mean values. *Mathematics of Computation* **42**: 193–194.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning* **1**: 81–106.