Theoretical Comparison between the Gini Index and Information Gain Criteria *

Laura E. Raileanu and Kilian Stoffel
University of Neuchâtel, Computer Science Departement
Pierre-à-Mazel 7, CH-2000 Neuchâtel
(Switzerland)
{Laura.Raileanu,Kilian.Stoffel}@unine.ch
Phone: ++41 32 718 1370 Fax: ++41 32 718 1231

Abstract

Knowledge Discovery in Databases (KDD) is an active and important research area with the promise for a high payoff in many business and scientific applications. One of the main tasks in KDD is classification. A particular efficient method for classification is decision tree induction. The selection of the attribute used at each node of the tree to split the data (split criterion) is crucial in order to correctly classify objects. Different split criteria were proposed in the literature (Information Gain, Gini Index, etc.). It is not obvious which of them will produce the best decision tree for a given data set. A large amount of empirical tests were conducted in order to answer this question. No conclusive results were found.

In this paper we introduce a formal methodology, which allows us to compare multiple split criteria. This permits us to present fundamental insights into the decision process. Furthermore, we are able to present a formal description of how to select between split criteria for a given data set. As an illustration we apply the methodology to the two widely used split criteria: Gini Index and Information Gain.

1 Introduction

Early work in the field of decision tree construction focused mainly on the definition and on the realization of classification systems. Such systems are described in: [9, 15, 2, 13, 12, 11, 16, 10]. All of them are using different measures of impurity / entropy / goodness to select the split attribute in order to construct the decision tree. Once a certain number of algorithms were defined, a lot of research was dedicated to compare them. This is a relatively hard task as the different systems evolved from different backgrounds: information theory, discriminant analysis, encoding techniques etc. These comparisons have been predominantly empirical. We briefly enumerate here some of the reported

*This work was supported by grant number 2100-056986.99 from the Swiss National Science Foundation.

experiments: [7, 1, 8, 12, 3, 19, 4, 5, 6]. In [18], the authors proposed a measure for the distance between the bias of two evaluation metrics and gave numerical approximations of it.

However, a thorough understanding of the behavior of the split functions demands an analytical and direct comparison between them, without using any other external measure. Our contribution in this paper is to introduce a formal methodology, which allows us to analytically compare multiple split criteria. This permits us to present fundamental insights into the decision process. Furthermore, we are able to present a formal description of how to select between split criteria for a given dataset. As an illustration we apply the methodology to the two widely used split criteria: Gini Index and Information Gain.

2 Notations

To realize a theoretical analysis we begin by introducing some notations and definitions. Let $\mathcal{L}$ be a learning sample, $\mathcal{L} = \{(x_1, c_1), \ldots, (x_{\|\mathcal{L}\|}, c_J)\}$. We denote by $\|\mathcal{L}\|$ the number of objects in $\mathcal{L}$. $\forall i \in \{1, \ldots, \|\mathcal{L}\|\}$, x_i is a measurement vector, $x_i \in \mathcal{X}$, $\mathcal{X}$ being the measurement space. $\forall i \in \{1, \ldots, J\}$, $c_i \in \mathcal{C}$, where $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ is the set of classes. The prior probability that an object belongs to a given class c_i, is given by $p(c_i) = \frac{\|c_i\|}{\|\mathcal{L}\|}$. Given a test T, with n possible outcomes, we denote by t_i the set of the objects in $\mathcal{L}$ having the outcome i. The probability that the test T has the outcome i is estimated by $p(t_i) = \frac{\|t_i\|}{\|\mathcal{L}\|}$. $\|c_i, t_j\|$ denotes the number of objects of $\mathcal{L}$ that lay in the class c_i and have the outcome j for the test T. The probability that an object lays in c_i and has the outcome j is given by $p(c_i, t_j) = \frac{\|c_i, t_j\|}{\|\mathcal{L}\|}$. The conditional probability, $p(c_i|t_j)$, that an object lays in the class c_i, under the condition that the test T has the outcome j, is estimated by $\frac{p(c_i, t_j)}{p(t_j)}$. Obviously we have: $\sum_{i=1}^{k} p(c_i) = 1$, $\sum_{i=1}^{k} p(c_i|t_j) = 1$, $\forall j \in \{1, \ldots, n\}$, $p(c_i), p(c_i|t_j), p(t_i) \in [0, 1]$ and $p(c_i|t_j) = \frac{p(c_i, t_j)}{p(t_j)}$ $\forall j \in \{1, \ldots, n\}$ and $\forall i \in \{1, \ldots, k\}$.

3 The Gini Index and Information Gain Criteria

In [2] the binary tree classifiers are constructed by repeatedly splitting subsets of $\mathcal{L}$ into two descendant subsets, beginning with $\mathcal{L}$ itself. To split $\mathcal{L}$ into smaller and smaller subsets we have to select the splits in such a way that the descendent subsets are always "purer" than their parents. Thus was introduced the "goodness of split" criterion, which is derived from the notion of an impurity function. An *impurity function* is a function ϕ defined on the set of all k-tuples of numbers $(p(c_1), p(c_2), \ldots, p(c_k))$ satisfying $p(c_i) \geq 0$, $\forall i \in \{1, \ldots, k\}$ and $\sum_{1=1}^{k} p(c_i) = 1$ with the following properties: a) ϕ achieves its maximum at the point $(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$; b) ϕ achieves its minimum at the points $(1, 0, \ldots, 0)$, $(0, 1, \ldots, 0), \ldots,$ $(0, 0, \ldots, 1)$; c) ϕ is a symmetric function of $(p(c_1), p(c_2), \ldots, p(c_k))$. Given an impurity function ϕ, *the impurity measure of any node* t is defined by: $i(t) = \phi(p(c_1|t), p(c_2|t), \ldots, p(c_k|t))$. If a split s in a node t divides all examples into two subsets t_1 and t_2 of proportions p_1 and p_2, *the decrease of impurity* is defined as: $\Delta i(s, t) = i(t) - p_1 i(t_1) - p_2 i(t_2)$. The *goodness of split* $\phi(s, t)$ is defined as $\Delta i(s, t)$. If a test T is used in a node t and this test is based on an attribute having n possible values, the expressions defined before are

generalized as follows: $i(t) = \phi(p(c_1|t), p(c_2|t), \ldots, p(c_k|t))$, $\Delta i(s,t) = i(t) - \sum_{j=1}^{n} p(t_j) i(t_j)$.

Breiman adopts in his work the *Gini diversity Index* which has the following form:

$$\phi(p(c_1|t), p(c_2|t), \ldots, p(c_k|t)) = \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} p(c_i|t)p(c_j|t) = 1 - \sum_{i=1}^{k} (p(c_i|t))^2 \quad (1)$$

In a node t an impurity function based on the Gini Index criterion assigns an example to a class c_i with the probability $p(c_i|t)$. The estimated probability that the item is actually in class j is $p(c_j|t)$. Therefore, the estimated probability of misclassification under this rule is the Gini Index: $i(t) = \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} p(c_i|t)p(c_j|t) = 1 - \sum_{j=1}^{k} (p(c_j|t))^2$. This function can also be interpreted in terms of variance. In a node t we assign to all examples belonging to class c_j the value 1, and to all other examples the value 0. The sample variance of these values is: $p(c_j|t)(1 - p(c_j|t))$. There are k classes, thus the corresponding variances are summed together: $i(t) = \sum_{j=1}^{k} p(c_j|t)(1 - p(c_j|t)) = 1 - \sum_{j=1}^{k} (p(c_j|t))^2$. Having a test T with n outcomes the goodness of the split is expressed using the Gini Index as follows:

$$gini(T) = 1 - \sum_{i=1}^{k} (p(c_i))^2 - \sum_{i=1}^{n} p(t_i) \sum_{j=1}^{k} p(c_j|t_i)(1 - p(c_j|t_i)) \quad (2)$$

The Gini Index criterion selects a test that maximizes this function.

The Information Gain function [12] has its origin in the information theory. It is based on the notion of entropy, which characterizes the impurity of an arbitrary set of examples. If we randomly select an example of a set and we announce that it belongs to the class c_i, then the probability of this message is equal to $p(c_i) = \frac{\|c_i\|}{\|\mathcal{L}\|}$, and the amount of information it conveys is $-log_2(p(c_i))$. The expected information provided by a message with respect to the class membership can be expressed as:

$$info(\mathcal{L}) = -\sum_{i=1}^{k} p(c_i) log_2(p(c_i)) \quad (3)$$

The quantity $info(\mathcal{L})$ measures the average amount of information needed to identify the class of an example in $\mathcal{L}$. This quantity is also known as the *entropy of the set* $\mathcal{L}$ relative to the k-wise classification. The logarithm is in base 2 because the entropy is a measure of the expected encoding length measured in bits. We will consider a similar measurement after $\mathcal{L}$ has been partitioned in accordance with the n outcomes of a test T. The expected information requirement is the weighted sum over the subsets: $info_T(\mathcal{L}) = \sum_{i=1}^{n} p(t_i) info(T_i)$. The information gained by partitioning $\mathcal{L}$ in accordance to the test T is measured by the quantity: $gain(T) = info(\mathcal{L}) - info_T(\mathcal{L})$. We can rewrite the Information Gain as:

$$gain(T) = -\sum_{i=1}^{k} p(c_i) \log_2(p(c_i)) + \sum_{i=1}^{n} p(t_i) \sum_{j=1}^{k} p(c_j|t_i) \log_2(p(c_j|t_i)) \quad (4)$$

The Information Gain criterion selects a test that maximizes the Information Gain function.

So, the selected test by these criteria, T^*, will satisfy: $gini(T^*) = \max_{\forall T\ possible\ test} gini(T)$ and $gain(T^*) = \max_{\forall T\ possible\ test} gain(T)$ respectively. Therefore, we have: $gini(T^*) \geq gini(T), \forall T\ possible\ test$ and $gain(T^*) \geq gain(T), \forall T\ possible\ test$.

In order to obtain a characterization of these two criteria and to compare them, we restraint them, without loss of generality, to the situation in which we have only two possible outcomes for the test T, $n = 2$, and two possible classes $k = 2$. Therefore, we have:

$$gini(T) = 1 - \sum_{i=1}^{2}(p(c_i))^2 - \sum_{i=1}^{2} p(t_i) \sum_{j=1}^{2} p(c_j|t_i)(1 - p(c_j|t_i)) \quad (5)$$

$$gain(T) = -\sum_{i=1}^{2} p(c_i)\log_2(p(c_i)) + \sum_{i=1}^{2} p(t_i) \sum_{j=1}^{2} p(c_j|t_i)\log_2(p(c_j|t_i)) \quad (6)$$

For simplicity we denote by: $x = p(c_1)$, $r = p(t_1)$, $p = p(c_1|t_1)$ and $q = p(c_1|t_2)$. We have: $1 - x = p(c_2)$, $1 - r = p(t_2)$, $1 - p = p(c_2|t_1)$ and $1 - q = p(c_2|t_2)$. Using these notations and some simple calculations we rewrite the Gini Index and the Information Gain functions as:

$$gini(T) = 2x(1-x) - 2rp(1-p) - 2(1-r)q(1-q) \quad (7)$$

$$\begin{aligned} gain(T) &= -x\log_2(x) - (1-x)\log_2(1-x) + r[p\log_2(p) + (1-p)\log_2(1-p)] \\ &\quad +(1-r)[q\log_2(q) + (1-q)\log_2(1-q)] \end{aligned} \quad (8)$$

where $x, p, q \in (0, 1)$ and $r \in [0, 1]$.

4 Theoretical Analysis of the Gini Index and Information Gain Criteria

Let us suppose we have two tests, T, T' (based on two different attributes) which are used to split a given node. Now we analyze if the Gini Index criterion and the Information Gain criterion will select the same test. If this is not the case, we would like to know under which conditions the two criteria select differently.

First we will write the Gini Index (Information Gain) functions for the tests T, T':

$$\begin{aligned} gini(T) &= 2x(1-x) - 2rp(1-p) - 2(1-r)q(1-q) \\ gini(T') &= 2x'(1-x') - 2r'p'(1-p') - 2(1-r')q'(1-q') \end{aligned} \quad (9)$$

$$\begin{aligned} gain(T) &= -x\log_2(x) - (1-x)\log_2(1-x) + r[p\log_2(p) + \\ &\quad +(1-p)\log_2(1-p)] + (1-r)[q\log_2(q) + (1-q)\log_2(1-q)] \\ gain(T') &= -x'\log_2(x') - (1-x')\log_2(1-x') + r'[p'\log_2(p') + \\ &\quad +(1-p')\log_2(1-p')] + (1-r')[q'\log_2(q') + (1-q')\log_2(1-q')] \end{aligned} \quad (10)$$

where $x, p, q, p', q' \in (0, 1)$ and $r, r' \in [0, 1]$.

We observe that $x = x'$ as $x = p(c_1) = \frac{\|c_1\|}{\|\mathcal{L}\|} = x'$. This probability remains constant, independently of the selected test. The number of examples belonging to the class c_1 and to the class c_2 respectively, remains constant, independently of the selected test, and therefore, the following relation holds:

$$r(p - q) + q = r'(p' - q') + q' \quad (11)$$

r relates to r', p, q, p', q' and, respectively r' relates to r, p, q, p', q' as follows:

$$r = \frac{r'(p' - q') + q' - q}{p - q},\ p \neq q, \qquad r' = \frac{r(p - q) + q - q'}{p' - q'},\ p' \neq q'. \tag{12}$$

The cases $p = q$, $p' = q'$, and $q = q'$ will be treated separately.

Furthermore, the following conditions must be satisfied:

$$r' \geq 0 \Leftrightarrow \frac{r(p - q) + q - q'}{p' - q'} \geq 0,\ p' \neq q' \qquad r \geq 0 \Leftrightarrow \frac{r'(p' - q') + q' - q}{p - q} \geq 0,\ p \neq q \tag{13}$$

$$r' \leq 1 \Leftrightarrow \frac{r(p - q) + q - p'}{p' - q'} \leq 1,\ p' \neq q' \qquad r \leq 1 \Leftrightarrow \frac{r'(p' - q') + q' - p}{p - q} \leq 1,\ p \neq q \tag{14}$$

$$p, q, p', q' \in [0, 1] \tag{15}$$

The difference between the Gini Index functions corresponding to T and T' can be written using (12) as:

$$\begin{aligned} &gini(T) - gini(T') = 2r'p'(1 - p') + 2(1 - r')q'(1 - q') - 2rp(1 - p) - 2(1 - r)q(1 - q) \\ &= 2(r'q'^2 - r'p'^2 + r'p' - r'q' - q'^2 + q') - 2(rq^2 - rp^2 + rp - rq - q^2 + q) \\ &= 2[r'(q' - p')(q' + p') + r(p - q)(p + q) + (q - q')(q + q')] \\ &= 2[r(p - q)(p + q - p' - q') + (q - q')(q - p')] \end{aligned}$$

where $p, q, r, p', q', r' \in [0, 1]$.

To simplify our expression, we introduce f_1:

$$f_1 = \frac{(q' - q)(q - p')}{(p - q)(p + q - p' - q')},\ p \neq q,\ p + q \neq p' + q' \tag{16}$$

If the difference between the Gini Index functions corresponding to the tests T, T' is positive, then the favorite test for the Gini Index criterion is T, otherwise the favorite test is T'. The same holds for the Information Gain functions.

The difference corresponding to the Information Gain functions is expressed as follows:

$$\begin{aligned} &gain(T) - gain(T') = r[p\log_2(p) + (1 - p)\log_2(1 - p)] + (1 - r)[q\log_2(q) + (1 - q)\log_2(1 - q)] - \\ &- r'[p'\log_2(p') + (1 - p')\log_2(1 - p')] + (1 - r')[q'\log_2(q') + (1 - q')\log_2(1 - q')] \end{aligned}$$

where $p, q, p', q' \in (0, 1)$ and $r, r' \in [0, 1]$.

To simplify this expression, we will use the function $f(x)$ to substitute $x\log_2(x) + (1 - x)\log_2(1 - x)$. $f : (0, 1) \to [-1, 0)$. It's derivative is negative on the interval $\left(0, \frac{1}{2}\right]$ and positive on the interval $\left[\frac{1}{2}, 1\right)$. It's second derivative is positive on $(0, 1)$. Thus, this function is monotonically decreasing on $\left(0, \frac{1}{2}\right]$ and monotonically increasing on $\left[\frac{1}{2}, 1\right)$. It is a strictly convex function. Using it and (12), the difference between the Information Gain functions corresponding

to the tests T, T' is rewritten as:

$$gain(T) - gain(T') = r(f(p) - f(q)) - r'(f(p') - f(q')) + f(q) - f(q')$$
$$= r(f(p) - f(q)) - \frac{r(p-q)+q-q'}{p'-q'}(f(p') - f(q')) + f(q) - f(q')$$
$$= r[(f(p) - f(q)) - \frac{p-q}{p'-q'}(f(p') - f(q'))] - \frac{q-q'}{p'-q'}(f(p') - f(q')) + f(q) - f(q')$$
$$= \frac{r}{p'-q'}[(f(p) - f(q))(p' - q') - (f(p') - f(q'))(p - q)] - \frac{1}{p'-q'}[(q - q')\cdot$$
$$\cdot(f(p') - f(q')) - (f(q) - f(q'))(p' - q')]$$
$$= \frac{1}{p'-q'}\{r[(f(p) - f(q))(p' - q') - (f(p') - f(q'))(q - p)] + (f(q) - f(q'))\cdot$$
$$\cdot(p' - q') - (f(p') - f(q'))(q - q')\}$$

Now we apply the Lagrange theorem to the function f on the intervals: $[p, q]$, $[p', q']$, and $[q, q']$. The function f is continuous on $[p, q]$, it's derivative f' exists and it is finite on $[p, q]$, so by the Lagrange theorem we have: $\exists\ x_1 \in (p, q),\ f'(x_1) = \frac{f(p)-f(q)}{p-q}$. For $[p', q']$ the theorem's conditions are also satisfied and therefore: $\exists\ x_2 \in (p', q'),\ f'(x_2) = \frac{f(p')-f(q')}{p'-q'}$ and similarly for $[q, q']$ we have: $\exists\ x_3 \in (q, q'),\ f'(x_3) = \frac{f(q)-f(q')}{q-q'}$.

We express the Information Gain difference as:

$$gain(T) - gain(T') = \frac{1}{p'-q'}\{r[(f'(x_1)(p - q)(p' - q') - f'(x_2)(p' - q')(q - q')] +$$
$$+f'(x_3)(q - q')(p' - q') - f'(x_2)(p' - q')(q - q')\}$$
$$= r[f'(x_1)(p - q) - f'(x_2)(p - q)] + f'(x_3)(q - q') - f'(x_2)(q - q')$$
$$= r(p - q)(f'(x_1) - f'(x_2)) + (q - q')(f'(x_3) - f'(x_2)) = rE_1 + E_2$$

where $p' \neq q'$, $E_1 = (p - q)(f'(x_1) - f'(x_2))$ and $E_2 = (q - q')(f'(x_3) - f'(x_2))$. We will establish the sign of this difference under the conditions (13), (14), $p \neq q$, $p' \neq q'$ and $q \neq q'$. We denote by f_2 the ratio:

$$f_2 = \frac{-E_2}{E_1} = \frac{(q - q')(f'(x_2) - f'(x_3))}{(q - p)(f'(x_2) - f'(x_1))} \tag{17}$$

The following proposition is used in our analysis to establish the order of the points x_1, x_2, x_3.

Proposition: If f is a strictly convex function defined on $(0, 1)$ and $0 < a < b < c < 1$ we have:

$$\frac{f(b) - f(a)}{b - a} < \frac{f(c) - f(a)}{c - a} < \frac{f(c) - f(b)}{c - b} \tag{18}$$

Proof: $a < b < c \Rightarrow b = \lambda a + (1 - \lambda)c$, with $\lambda \in (0, 1)$. $f(b) = f(\lambda a + (1 - \lambda)c) < \lambda f(a) + (1 - \lambda)f(c)$ by the strictly convexity of f. We have $f(b) - f(a) < (1-\lambda)(f(c) - f(a))$. So $\frac{f(b)-f(a)}{b-a} < \frac{(1-\lambda)(f(c)-f(a))}{(1-\lambda)(c-a)} = \frac{f(c)-f(a)}{c-a}$ $(*)$. We have $f(b) - f(c) = f(\lambda a + (1 - \lambda)c) - f(c) < \lambda(f(a) - f(c))$. So $\frac{f(b)-f(c)}{b-c} > \frac{\lambda(f(a)-f(c))}{\lambda(a-c)} = \frac{f(a)-f(c)}{a-c}$ $(**)$. The proposition results from $(*)$ and $(**)$.□

As $r, r' \in [0, 1]$ and (12) must be satisfied, the terms $p' - q'$, $q' - q$, and $q - p$ can not be simultaneously positive or simultaneously negative, consequently, two terms are negative and one is positive or one term is negative and two terms are positive. Thus, the characterization of the Gini Index and Information Gain functions will be done taking into account only the six possible cases: 1) $p' - q' > 0$, $q - p > 0$, $q' - q < 0$, 2) $p' - q' > 0$, $q - p < 0$, $q' - q > 0$, 3) $p' - q' < 0$, $q - p > 0$, $q' - q > 0$, 4) $p' - q' < 0$, $q - p < 0$, $q' - q > 0$, 5) $p' - q' < 0$, $q - p > 0$, $q' - q < 0$, and 6) $p' - q' > 0$, $q - p < 0$, $q' - q < 0$.

5 Results

In this section, we present the intervals of coincidence/non-coincidence in the choice of the split attribute for the Gini Index and Information Gain criteria. The sign of the differences of the Gini Index functions corresponding to two tests T, T' and of the Information Gain functions are established for the six possible situations. We will present in the following the details for one case as an illustration. The complete analysis can be found in [14]. If the sign of the difference of the Gini Index functions $gini(T) - gini(T')$ is the same as the sign of the difference of the Information Gain functions $gain(T) - gain(T')$, then the two split criteria select the same attribute to split on, otherwise they select different attributes to split on.

Case 1: $\mathbf{p' - q' > 0}$, $\mathbf{q - p > 0}$, $\mathbf{q' - q < 0}$. This case can be subdivided into following subcases:

(a) $0 < p < q' < q < p' < 1$ (b) $0 < p < q' < p' < q < 1$ (c) $0 < q' < p < q < p' < 1$

(d) $0 < q' < p < p' < q < 1$ (e) $0 < q' < p' < p < q < 1$

Case 1.(a): $\mathbf{0 < p < q' < q < p' < 1}$

Proof: It is easy to show that $f_1 \in [0, 1]$. $q' - q < 0$, $q - p' < 0$, $p - q < 0$ and $p + q - p' - q' < 0$. $f_1 - 1 = \frac{(q'-p)(p-p')}{(p-q)(p+q-p'-q')} < 0$ as $q' - p > 0$, $p - p' < 0$, $p - q < 0$, and $p + q - p' - q' < 0$. For r and r' we must assure that (13), (14) are satisfied. (14) are satisfied. But to verify that (13) are satisfied, it is necessary that $r \leq \frac{q'-q}{p-q}$ and $r' \leq \frac{q-q'}{p'-q'}$. Both ratios: $\frac{q'-q}{p-q}$, $\frac{q-q'}{p'-q'}$ are positive and smaller than 1, so we can conclude that for this case we have: $r \in \left[0, \frac{q'-q}{p-q}\right]$ and $r' \in \left[0, \frac{q-q'}{p'-q'}\right]$. In addition we can easily show that $f_1 < \frac{q'-q}{p-q}$.

Knowing now the position of r and f_1 relative to $\frac{q'-q}{p-q}$ we can establish the sign of the difference between $gini(T)$ and $gini(T')$. For $r \in [0, f_1]$ we have $gini(T) - gini(T') \leq 0$ and for $r \in \left[f_1, \frac{q'-q}{p-q}\right]$ we have $gini(T) - gini(T') \geq 0$.

To evaluate the difference between the $gain(T)$ and $gain(T')$ we proceed in the same way. The conditions obtained for r and r' remain valid. We must find this time the position of f_2 and of r. First, we must establish the order of x_1, x_2, x_3. These points can be ordered by considering all the possible permutations of them. Applying the proposition (18) to the points $p < q < p'$, $p < q' < q$, $p < q' < p'$, and $q' < q < p'$ we find that $f'(x_1) < f'(x_3) < f'(x_2)$. And, using that f' is strictly monotonically increasing (its derivative, f'', is positive), we conclude that we have to analyze only the case $x_1 < x_3 < x_2$. The other cases contradict the monotonicity of f'. Now, it is easy to show that $E_1 \geq 0$, $E_2 \leq 0$ and $f_2 \in \left[0, \frac{q'-q}{p-q}\right]$. We have $f_2 \leq \frac{q'-q}{p-q}$ as: $f_2 \leq \frac{q'-q}{p-q} \Leftrightarrow \frac{f'(x_2)-f'(x_3)}{f'(x_2)-f'(x_1)} \leq 1 \Leftrightarrow f'(x_3) \geq f'(x_1)$. So, if $r \in [0, f_2]$ we have $gain(T) - gain(T') \leq 0$, and if $r \in \left[f_2, \frac{q'-q}{p-q}\right]$ we have $gain(T) - gain(T') \geq 0$.

In conclusion, for $0 < p < q' < q < p' < 1$ we have: $r \in \left[0, \frac{q'-q}{p-q}\right]$, $r' \in \left[0, \frac{q-q'}{p'-q'}\right]$ and $f_1, f_2 \in \left[0, \frac{q'-q}{p-q}\right]$. If $r \in [0, \min\{f_1, f_2\}]$ the same split is selected. If $r \in (\min\{f_1, f_2\}, \max\{f_1, f_2\})$ different splits are selected. If $r \in \left[\max\{f_1, f_2\}, \frac{q'-q}{p-q}\right]$ the same split is selected.

Case 1.(b): $\mathbf{0 < p < q' < p' < q < 1}$

Proof: To establish the position of r and r' we use the conditions (13) and (14) as we did before, and we obtain: $r \in \left[\frac{p'-q}{p-q}, \frac{q'-q}{p-q}\right]$ and $r' \in [0, 1]$. If $p + q - p' - q' \leq 0 \Rightarrow gini(T) - gini(T') \geq 0$. If $p + q - p' - q' \geq 0 \Rightarrow f_1 \geq$

$1 \Rightarrow r \leq f_1 \Rightarrow gini(T) - gini(T') \geq 0$. For $r \in [\frac{p'-q}{p-q}, \frac{q'-q}{p-q}]$ and $r' \in [0,1]$ we have $gini(T) - gini(T') \geq 0$.

To evaluate the difference between the $gain(T)$ and $gain(T')$ we proceed in the same way. The conditions obtained for r and r' remain valid. The points x_1, x_2, x_3 will be ordered as in the previous case. Applying proposition (18) to the points $p < q' < p'$, $p < q' < q$, $q' < p' < q$, and $p < p' < q$ and using that f' is strictly monotonically increasing, we conclude that we have only two possible cases: $x_1 < x_2 < x_3$ and $x_2 < x_1 < x_3$. In the case $x_1 < x_2 < x_3$ we have: $E_1 \geq 0$ and $E_2 \geq 0$. So $gain(T) - gain(T') \geq 0$. In the case $x_2 < x_1 < x_3$ we have $E_1 \leq 0$, $E_2 \geq 0$, and $f_2 \geq \frac{q'-q}{p-q}$. So for $r \in \left[\frac{p'-q}{p-q}, \frac{q'-q}{p-q}\right]$ we have $gain(T) - gain(T') \geq 0$.

Proof for $f_2 \geq \frac{q'-q}{p-q}$: $f_2 \geq \frac{q'-q}{p-q} \Leftrightarrow \frac{f'(x_2)-f'(x_3)}{f'(x_2)-f'(x_1)} \geq 1 \Leftrightarrow f'(x_3) \geq f'(x_1)$ which is true. □

In conclusion, for $0 < p < q' < p' < q < 1$ we have: $r \in \left[\frac{p'-q}{p-q}, \frac{q'-q}{p-q}\right]$, $r' \in [0,1]$, and the behavior of the two split functions is identical, both are choosing T as split.

Case 1.(c): $0 < q' < p < q < p' < 1$

Proof: We have: $r \in [0,1]$ and $r' \in \left[\frac{p-q'}{p'-q'}, \frac{q-q'}{p'-q'}\right]$. If $p + q - p' - q' \leq 0 \Rightarrow gini(T) - gini(T') \leq 0$. If $p + q - p' - q' \geq 0 \Rightarrow f_1 \geq 1 \Rightarrow r \leq f_1 \Rightarrow gini(T) - gini(T') \leq 0$. For $r \in [0,1]$ and $r' \in \left[\frac{p-q'}{p'-q'}, \frac{q-q'}{p'-q'}\right]$ we have $gini(T) - gini(T') \leq 0$.

Applying proposition (18) to the points $q' < p < q$, $q' < p < p'$, $q' < q < p'$, and $p < q < p'$ and, using the fact that f' is strictly monotonically increasing, we conclude that we have only the following cases to analyze: $x_3 < x_1 < x_2$ and $x_3 < x_2 < x_1$. If $x_3 < x_1 < x_2$ we have: $E_1 \geq 0$, $E_2 \leq 0$, and $f_2 > 1$, and therefore, $gain(T) - gain(T') < 0$.

Proof for $f_2 > 1$: $f_2 > 1 \Leftrightarrow \frac{f'(x_2)-f'(x_3)}{f'(x_2)-f'(x_1)} > \frac{q-p}{q-q'}$ and this is true as $\frac{f'(x_2)-f'(x_3)}{f'(x_2)-f'(x_1)} \geq 1 \Leftrightarrow f'(x_3) \leq f'(x_1)$ and $\frac{q-p}{q-q'} < 1 \Leftrightarrow p > q'$. □

For the other situation $x_3 < x_2 < x_1$ we have: $E_1 \leq 0$, $E_2 \leq 0$. So $gain(T) - gain(T') \leq 0$.

In conclusion, for $0 < q' < p < q < p' < 1$ we have: $r \in [0,1]$, $r' \in \left[\frac{p-q'}{p'-q'}, \frac{q-q'}{p'-q'}\right]$, and the behavior of the two split functions is identical, both are choosing T' as split.

Case 1.(d): $0 < q' < p < p' < q < 1$

Proof: Analogously we obtain: $r \in \left[\frac{p'-q}{p-q}, 1\right]$, $r' \in \left[\frac{p-q'}{p'-q'}, 1\right]$ and $f_1 \in \left[\frac{p'-q}{p-q}, 1\right]$. If $r \in \left[\frac{p'-q}{p-q}, f_1\right] \Rightarrow gini(T) - gini(T') \geq 0$. If $r \in [f_1, 1] \Rightarrow gini(T) - gini(T') \leq 0$.

Applying the proposition (18) to the points $q' < p < p'$, $q' < p < q$, $q' < p' < q$, and $p < p' < q$ and, using that f' is strictly monotonically increasing we conclude that we have only the case $x_2 < x_3 < x_1$ to analyze. As $x_2 < x_3 < x_1$ we have: $E_1 \leq 0$, $E_2 \geq 0$, and $f_2 \in \left(\frac{p'-q}{p-q}, 1\right)$. For $r \in \left[\frac{p'-q}{p-q}, f_2\right]$ we have $gain(T) - gain(T') \geq 0$ and for $r \in [f_2, 1]$ we have $gain(T) - gain(T') \leq 0$.

Proof for $f_2 > \frac{p'-q}{p-q}$: $f_2 > \frac{p'-q}{p-q} \Leftrightarrow (f'(x_3) - f'(x_2))(q - q') > (f'(x_1) - f'(x_2))(q - p') \Leftrightarrow f'(x_3)(q - q') + f'(x_2)(q' - p') + f'(x_1)(p' - q) > 0 \Leftrightarrow \frac{f(q)-f(q')}{q-q'}(q - q') + \frac{f(q')-f(p')}{q'-p'}(q' - p') + \frac{f(q)-f(p)}{q-p}(p' - q) > 0 \Leftrightarrow$

$f(q)(p'-p)-f(p)(p'-q)-f(p')(q-p)>0$ $(*)$. As $p<p'<q$, $p'=\alpha p+(1-\alpha)q$ with $\alpha\in(0,1)$, so we have $(*)\Leftrightarrow f(q)(1-\alpha)+f(p)\alpha>f(\alpha p+(1-\alpha)q)$ which is true by the strict convexity of f. □

Proof for $f_2<1$: $f_2<1\Leftrightarrow(f'(x_2)-f'(x_3))(q-q')>(f'(x_2)-f'(x_1))(q-p)\Leftrightarrow f'(x_3)(q'-q)+f'(x_2)(p-q')+f'(x_1)(q-p)>0\Leftrightarrow\frac{f(q')-f(q)}{q'-q}(q'-q)+\frac{f(p')-f(q')}{p'-q'}(p-q')+\frac{f(q)-f(p)}{q-p}(q-p)>0\Leftrightarrow f(q')(p'-p)-f(p)(p'-q')+f(p')(p-q')>0$ $(**)$. As $q'<p<p'$, $p=\alpha q'+(1-\alpha)p'$ with $\alpha\in(0,1)$, so we have $(**)$ $\Leftrightarrow f(p')(1-\alpha)+f(q')\alpha>f(\alpha q'+(1-\alpha)p')$ which is true by the strict convexity of f. □

In conclusion, for $0<q'<p<p'<q<1$ we have $r\in\left[\frac{p'-q}{p-q},1\right]$ and $r'\in\left[\frac{p-q'}{p'-q'},1\right]$. For $r\in\left[\frac{p'-q}{p-q},\min\{f_1,f_2\}\right]$ we have the same split, for $r\in(\min\{f_1,f_2\},\max\{f_1,f_2\})$ we have different splits, and for $r\in[\max\{f_1,f_2\},1]$ we have the same split.

Case 1.(e): $0<q'<p'<p<q<1$

Proof: This case is dropped as it contradicts the conditions (14).

The other possible cases listed are treated in the same manner as the first one. Each of the remaining cases is divided in several sub-cases by taking into account the position of p,q,p',q'. The domains of r,r',f_1,f_2 are established for each sub-case following an identical path as for the first case. The complete detailed analysis can be found in [14]. The results obtained for the six cases identified can be resumed in the following way. For the case 1) we obtained two situations in which the two split criteria are selecting different tests; by symmetry we obtain for the case 4) two such situations. Cases 2) and 5) are similar (also by the symmetry) and for each of them we obtain one situation in which the selection of test is done differently by the two criteria. Finally, cases 3) and 6) are symmetric, and for each of them we obtain a situation of different selection.

By this formal analysis, we were able to study the behavior of the Gini Index and Information Gain, to give an exact mathematical description of the situations when they are choosing the same test to split on and when not. This allows us, without constructing decision trees, to decide for a given database if the Gini Index criterion and the Information Gain criterion select the same split attribute.

In order to compare the two split functions in a general way, we used the obtained results to compute the frequency of agreement or disagreement of the two split functions. In a sequence of tests, we calculated for all considered sizes of databases the number of cases of disagreement. It was never higher than 2%. This explains why most empirical studies concluded that there is no significant difference between the two criteria. Of course this does not exclude that for some specific databases there might be an important difference.

6 Conclusions and Future Work

In this paper, we presented a formal comparison of the behavior of two of the most popular split functions, namely the Gini Index function and the Information Gain function. The situations where the two split functions agree/disagree on the selected split were mathematically characterized. Based on these characterizations we were able to analyze

the frequency of agreement/disagremment of the Gini Index function and the Information Gain function. We found that they disagree only in 2%, which explains why most previously published empirical results concluded that it is not possible to decide which one of the two tests to prefer. Moreover we would like to emphasize that the methodology introduced in this paper is not limited to the two analyzed split criteria. We used it successfully to formalize and compare other split criteria. Based on the gained deeper insights on the split process we are currently working on a system, which will select the optimal criterion based on a user defined optimality criterion. Preliminary results can be found in [17].

References

[1] A. Babic, E. Krusinska, and J. E. Stromberg. Extraction of diagnostic rules using recursive partitioning systems: A comparison of two approches. *Artificial Intelligence in Medicine*, 20(5):373–387, October 1992.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Wadsworth International Group, 1984.

[3] Lopez de Mantaras. A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6(1):81–92, 1991.

[4] J. Gama and P. Brazdil. Characterization of classification algorithms. In C. Pinto-Ferreira and N. Mamede, editors, *EPIA-95: Progress in Artificial Intelligence, 7th Portuguese Conference on Artificial Intelligence*, pages 189–200. Springer Verlag, 1995.

[5] Igor Kononenko. On biases in estimating multi-valued attributes. In Chris Mellish, editor, *IJCAI-95: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1034–1040, Montreal, Canada, August 1995. Morgan Kaufmann Publishers Inc, San Mateo, CA.

[6] Tjen-Sien Lim, Wey-Yin Loh, and Yu-Shan Shih. A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms. *Machine Learning*, 1999.

[7] John Mingers. An empirical comparison of selection measures for decision tree induction. *Machine Learning*, 3:319–342, 1989.

[8] Masashiro Miyakawa. Criteria for selecting a variable in the construction of efficient decision trees. *IEEE Transactions on Computers*, 35(1):133–141, January 1929.

[9] B. M. Moret. Decision trees and diagrams. *Computing Surveys*, 14(4):593–623, 1982.

[10] Kolluru Venkata Sreerama Murthy. *On Growing Better Decision Trees from Data*. PhD thesis, The John Hopkins University, Baltimore, Maryland, 1995.

[11] G. Pagallo. *Adaptive Decision Tree Algorithms for Learning from Examples*. PhD thesis, University of California, Santa Cruz, 1990.

[12] J. R. Quinlan. *C4.5 Programs for machine learning*. Morgan Kaufmann Publishers, 1993.

[13] John Ross Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, (27):221–234, 1987.

[14] Laura E. Raileanu. Theoretical comparison between the gini index and information gain functions. Technical report, Faculté de droit et Sciences Economiques, Université de Neuchâtel, 2000.

[15] S. R. Safavin and D. Langrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics*, 21(3):660–674, 1991.

[16] M. Sahami. Learning non-linearly separable boolean functions with linear threshold unit trees and madaline-style networks. In AAAI Press, editor, *Proceedings of the Eleventh National Conference on Artificial Inteligence*, pages 335–341, 1993.

[17] Kilian Stoffel and Laura E. Raileanu. Selecting optimal split-functions for large datasets. In *Research and Development in Intelligent Systems XVII*, BCS Conference Series, 2000.

[18] Ricardo Vilalta and Daniel Oblinger. A quantification of distance-bias between evaluation metrics in classification. In *Proceedings of the 17th International Conference on Machine Learning*. Stanford University, 2000.

[19] Allan P. White and Wei Zhang Liu. Bias il information-based measures in decision tree induction. *Machine Learning*, 15(3):321–328, June 1997.