Annals of Mathematics and Artificial Intelligence **41:** 77–93, 2004.

Theoretical comparison between the Gini Index and Information Gain criteria *

Laura Elena Raileanu and Kilian Stoffel

University of Neuchâtel, Computer Science Department, Pierre-à-Mazel 7, CH-2000 Neuchâtel, Switzerland

E-mail: {laura.raileanu,kilian.stoffel}@unine.ch

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 two widely used split criteria: Gini Index and Information Gain.

Keywords: decision trees, classification, Gini Index, Information Gain, theoretical comparison

AMS subject classification: 68T15, 68T20, 68T30, 68T35

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 [4,12–16,18,19]. All of them use 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 difficult task as the systems evolved from different backgrounds: information theory, discriminant analysis, encoding techniques, etc. These comparisons have been predominantly empirical. Baker and Jain [2] reported experiments comparing eleven feature evaluation criteria and concluded that the feature rankings induced by various rules are very similar. Several feature evaluation criteria are compared using simulated data by Ben-Bassat [3], on a sequential, multi-class clas-

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

sification problem. The conclusions are that no feature selection rule is consistently superior to the others, and that no specific strategy for alternating different rules is significantly more effective. Mingers [10] compared several attribute selection criteria, and concluded that the tree quality does not seem to depend on the specific criterion used. Babic [1] compared ID3 and CART for two clinical diagnosis problems. Miyakawa [11] compared three activity-based measures, both analytically and empirically. Several researchers pointed out that Information Gain is biased towards attributes with a large number of possible values. Mingers [9] compared Information Gain and χ^2-statistic for growing the tree as well as for stop splitting. He concluded that χ^2-corrected Information Gain's bias towards multi-valued attributes. Quinlan [16] suggested Gain Ratio as a remedy for the bias of Information Gain. Mantaras [5] argued that Gain Ratio had its own set of problems, and suggested information theory based distance between partitions for tree constructions. White and Liu [22] present experiments to conclude that Information Gain, Gain Ratio and Mantara's measure are worse than a χ^2-based statistical measure, in terms of their bias towards multiple-valued attributes. Gama [6] in Esprit Project 5170 StatLog (1991–1994) tried to predict the error rate of a particular classification algorithm and he indicated that no single method can be considered better than the others. About twenty different algorithms were evaluated on more than twenty different data sets. Kononenko [7] pointed out that Minimum Description Length based feature evaluation criteria have the least bias towards multi-valued attributes. In [8] twenty-two decision tree and two neural network algorithms are compared in terms of classification accuracy, training time, and number of leaves. In [20] Gini Index, Information Gain, and the new family of split functions are tested on 9000 data sets of different sizes (from 200 to 20 000 tuples). In [21], 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 two widely used split criteria: Gini Index and Information Gain.

The outline of this paper is the following: we introduce the notation and definitions required to present the classical split criteria which are central to construction of decision trees: the Gini Index and the Information Gain criteria. After, we give the description of the theoretical analysis of the Gini Index and Information Gain and we present the results obtained. Finally, we present some future work and the conclusions.

2. Notation

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. $\forall i \in \{1,\dots,J\}$, c_i represents the class i, and $c_i \in \mathcal{C}$, where $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$ is the set of classes. The prior probability that an object belongs to a given class c_i, is given by $p(c_i) = \frac{\|c_i\|}{\|\mathcal{L}\|}$. The components of the vectors x_i can be viewed as attributes and a test is based on one of these attributes. Given a test T (based on a single attribute), 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 are in the class c_i and have the outcome j for the test T. The probability that an object is 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 is 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: $\sum_{i=1}^{k} p(c_i) = 1$, $\sum_{i=1}^{k} p(c_i|t_j) = 1$ $\forall j \in \{1,\dots,n\}$, $p(c_i)$, $p(c_i|t_j)$, $p(t_i) \in [0,1]$ and $p(c_i|t_j) = \frac{p(c_i,t_j)}{p(t_j)}$ $\forall j \in \{1,\dots,n\}$ and $\forall i \in \{1,\dots,k\}$.

3. The Gini Index and Information Gain criteria

The objects are classified by decision trees which sort them down from the *root* to some leaf node, which provides the classification (the class) of each object. An decision tree contains zero or more *internal nodes* and one or more *leaf nodes*. The internal nodes have two or more *child nodes*. Each nonterminal node contains *a split* which specifies *a test* based on a single attribute, and each branch descending from that node corresponds to one of the possible values for this attribute. Each leaf node has its class label. A leaf node is said to be *pure* if all of its training examples are belonging to the same class.

Thus, an example is classified by starting with the root node, testing the attribute corresponding to the root node, then moving down the tree branch corresponding to the value of the attribute in the given example. This process continues until the example reaches a leaf node.

In [4] 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 $(p(c_1), p(c_2), \dots, p(c_k))$ satisfying $p(c_i) \geqslant 0$ $\forall i \in \{1,\dots,k\}$ and $\sum_{1=1}^{k} p(c_i) = 1$ with the following properties:

(a) ϕ achieves its maximum at the point $(\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k})$;

(b) ϕ achieves its minimum at the points $(1,0,\dots,0)$, $(0,1,\dots,0)$, $\dots$, $(0,0,\dots,1)$;

(c) ϕ is a symmetric function of $(p(c_1), p(c_2), \dots, p(c_k))$.

Given an impurity function ϕ, *the impurity measure of any node t* is defined by

$$i(t) = \phi\bigl(p(c_1|t), p(c_2|t), \dots, p(c_k|t)\bigr).$$

If a split s in a node t divides all examples into two subsets t_L and t_R of proportions p_L and p_R, *the decrease of impurity* is defined as

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R).$$

The *goodness of split s* in node t, $\phi(s,t)$, 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:

$$i(t) = \phi\big(p(c_1|t), p(c_2|t), \dots, p(c_k|t)\big),$$
$$\Delta i(s,t) = i(t) - \sum_{j=1}^{n} p(t_j) i(t_j).$$

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

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

In a node t, an impurity function based on the Gini Index criterion assigns a training 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:

$$i(t) = \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} p(c_i|t)p(c_j|t) = 1 - \sum_{j=1}^{k}\big(p(c_j|t)\big)^2.$$

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:

$$i(t) = \sum_{j=1}^{k} p(c_j|t)\big(1 - p(c_j|t)\big) = 1 - \sum_{j=1}^{k}\big(p(c_j|t)\big)^2.$$

Having a test T with n outcomes the goodness of the split is expressed using the Gini Index as follows:

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

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

The Information Gain function [16] has its origin in 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 from 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

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

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

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

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

So, the selected test by these criteria, T^*, will satisfy:

$$gini\bigl(T^*\bigr) = \max_{\forall\ possible\ test\ T} gini(T)$$

and

$$gain\bigl(T^*\bigr) = \max_{\forall\ possible\ test\ T} gain(T),$$

respectively. Therefore, we have: $gini(T^*) \geqslant gini(T)\ \forall\, possible\ test\ T$ and $gain(T^*) \geqslant gain(T)\ \forall\, possible\ test\ T$.

In order to obtain a characterization of these two criteria and to compare them, we restrict 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:

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

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

For simplicity we denote: $x = p(c_1)$, $r = p(t_1)$, $p = p(c_1|t_1)$ 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

$$gini(T) = 2x(1-x) - 2rp(1-p) - 2(1-r)q(1-q), \tag{7}$$

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

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

4. Theoretical analysis of the Gini Index and Information Gain criteria

In this section we give the description of the theoretical analysis of the Gini Index and Information Gain. 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':

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

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

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:

$$r(p-q) + q = r'(p'-q') + q'. \tag{11}$$

The statement above holds since:

$$r(p-q) + q = rp + q(1-r)$$

$$\begin{aligned}
&= p(t_1)\frac{p(c_1,t_1)}{p(t_1)} + \frac{p(c_1,t_2)}{p(t_2)}p(t_2)\\
&= p(c_1,t_1) + p(c_1,t_2) = p(c_1)
\end{aligned}$$

and

$$\begin{aligned}
r'\bigl(p'-q'\bigr)+q' &= r'p'+q'\bigl(1-r'\bigr)\\
&= p\bigl(t_1'\bigr)\frac{p(c_1,t_1')}{p(t_1')} + \frac{p(c_1,t_2')}{p(t_2')}p\bigl(t_2'\bigr)\\
&= p\bigl(c_1,t_1'\bigr) + p\bigl(c_1,t_2'\bigr) = p(c_1).
\end{aligned}$$

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

$$\begin{aligned}
r &= \frac{r'(p'-q')+q'-q}{p-q}, \quad p \neq q,\\
r' &= \frac{r(p-q)+q-q'}{p'-q'}, \quad p' \neq q'.
\end{aligned} \tag{12}$$

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

Furthermore, using (12), the following conditions must be satisfied:

$$r' \geqslant 0 \iff \frac{r(p-q)+q-q'}{p'-q'} \geqslant 0, \quad p' \neq q', \tag{13}$$

$$r \geqslant 0 \iff \frac{r'(p'-q')+q'-q}{p-q} \geqslant 0, \quad p \neq q, \tag{14}$$

$$r' \leqslant 1 \iff r'-1 \leqslant 0 \iff \frac{r(p-q)+q-p'}{p'-q'} \leqslant 0, \quad p' \neq q', \tag{15}$$

$$r \leqslant 1 \iff r-1 \leqslant 0 \iff \frac{r'(p'-q')+q'-p}{p-q} \leqslant 0, \quad p \neq q, \tag{16}$$

$$p,q,p',q' \in [0,1]. \tag{17}$$

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

$$\begin{aligned}
&gini(T) - gini\bigl(T'\bigr)\\
&\quad = 2r'p'\bigl(1-p'\bigr) + 2\bigl(1-r'\bigr)q'\bigl(1-q'\bigr) - 2rp(1-p) - 2(1-r)q(1-q)\\
&\quad = 2\bigl(r'q'^2 - r'p'^2 + r'p' - r'q' - q'^2 + q'\bigr)\\
&\qquad - 2\bigl(rq^2 - rp^2 + rp - rq - q^2 + q\bigr)\\
&\quad = 2\bigl[r'\bigl(q'-p'\bigr)\bigl(q'+p'\bigr) + r(p-q)(p+q) + \bigl(q-q'\bigr)\bigl(q+q'\bigr)\bigr]\\
&\quad = 2\bigl[r(p-q)\bigl(p+q-p'-q'\bigr) + \bigl(q-q'\bigr)\bigl(q-p'\bigr)\bigr],
\end{aligned} \tag{18}$$

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

To simplify the obtained expression, we introduce f_1:

$$f_1 = \frac{(q' - q)(q - p')}{(p - q)(p + q - p' - q')}, \quad p \neq q, \ p + q \neq p' + q'. \tag{19}$$

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:

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

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

To simplify this expression, we will use the function $f(x) = x\log_2(x) + (1 - x)\log_2(1 - x)$, $f : (0, 1) \to [-1, 0)$. It's derivative is negative on the interval $(0, \frac{1}{2}]$ and positive on the interval $[\frac{1}{2}, 1)$. It's second derivative is positive on $(0, 1)$. Thus, this function is monotonically decreasing on $(0, \frac{1}{2}]$ and monotonically increasing on $[\frac{1}{2}, 1)$. It is a strictly convex function. Using the function f and (12), the difference between the Information Gain functions corresponding to the tests T, T' is rewritten as

$$\begin{aligned} & gain(T) - gain(T') \\ & \quad = r\big(f(p) - f(q)\big) - r'\big(f(p') - f(q')\big) + f(q) - f(q') \\ & \quad = r\big(f(p) - f(q)\big) - \frac{r(p - q) + q - q'}{p' - q'}\big(f(p') - f(q')\big) + f(q) - f(q') \\ & \quad = r\left[\big(f(p) - f(q)\big) - \frac{p - q}{p' - q'}\big(f(p') - f(q')\big)\right] \\ & \qquad - \frac{q - q'}{p' - q'}\big(f(p') - f(q')\big) + f(q) - f(q') \\ & \quad = \frac{r}{p' - q'}\big[\big(f(p) - f(q)\big)(p' - q') - \big(f(p') - f(q')\big)(p - q)\big] \\ & \qquad - \frac{1}{p' - q'}\big[(q - q')\big(f(p') - f(q')\big) - \big(f(q) - f(q')\big)(p' - q')\big] \\ & \quad = \frac{1}{p' - q'}\big\{r\big[\big(f(p) - f(q)\big)(p' - q') - \big(f(p') - f(q')\big)(q - p)\big] \\ & \qquad + \big(f(q) - f(q')\big)(p' - q') - \big(f(p') - f(q')\big)(q - q')\big\}. \end{aligned}$$

Now we apply the Lagrange theorem (also known as the Mean value 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:

$$\exists x_1 \in (p, q)\ f'(x_1) = \frac{f(p) - f(q)}{p - q}. \tag{20}$$

For $[p', q']$ the theorem's conditions are also satisfied and therefore:

$$\exists x_2 \in \left(p', q'\right)\ f'(x_2) = \frac{f(p') - f(q')}{p' - q'} \tag{21}$$

and similarly for $[q, q']$ we have:

$$\exists x_3 \in \left(q, q'\right)\ f'(x_3) = \frac{f(q) - f(q')}{q - q'}. \tag{22}$$

We express the Information Gain difference as

$$\begin{aligned}
&gain(T) - gain\left(T'\right) \\
&\quad = \frac{1}{p' - q'}\Big\{r\big[f'(x_1)(p - q)\left(p' - q'\right) - f'(x_2)\left(p' - q'\right)\left(q - q'\right)\big] \\
&\qquad\qquad + f'(x_3)\left(q - q'\right)\left(p' - q'\right) - f'(x_2)\left(p' - q'\right)\left(q - q'\right)\Big\} \\
&\quad = r\big[f'(x_1)(p - q) - f'(x_2)(p - q)\big] + f'(x_3)\left(q - q'\right) - f'(x_2)\left(q - q'\right) \\
&\quad = r(p - q)\big(f'(x_1) - f'(x_2)\big) + \left(q - q'\right)\big(f'(x_3) - f'(x_2)\big) \\
&\quad = rE_1 + E_2,
\end{aligned} \tag{23}$$

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), (15), $p \neq q$, $p' \neq q'$ and $q \neq q'$.

We denote by f_2 the ratio:

$$f_2 = \frac{-E_2}{E_1} = \frac{(q - q')(f'(x_2) - f'(x_3))}{(q - p)(f'(x_2) - f'(x_1))}. \tag{24}$$

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 $0 < a < b < c < 1$, then we have:

$$\frac{f(b) - f(a)}{b - a} < \frac{f(c) - f(a)}{c - a} < \frac{f(c) - f(b)}{c - b}. \tag{25}$$

Proof. If $a < b < c$ then $b = \lambda a + (1 - \lambda)c$, with $\lambda \in (0, 1)$ and

$$f(b) = f\big(\lambda a + (1 - \lambda)c\big) < \lambda f(a) + (1 - \lambda)f(c)$$

by the strictly convexity of f. We have

$$f(b) - f(a) < (1 - \lambda)\big(f(c) - f(a)\big).$$

So

$$\frac{f(b)-f(a)}{b-a} < \frac{(1-\lambda)(f(c)-f(a))}{(1-\lambda)(c-a)} = \frac{f(c)-f(a)}{c-a}. \qquad (*)$$

We have using the strictly convexity of f:

$$f(b)-f(c) = f\big(\lambda a + (1-\lambda)c\big) - f(c) < \lambda\big(f(a)-f(c)\big).$$

So

$$\frac{f(b)-f(c)}{b-c} > \frac{\lambda(f(a)-f(c))}{\lambda(a-c)} = \frac{f(a)-f(c)}{a-c}. \qquad (**)$$

The proposition results from $(*)$ and $(**)$. □

As $r, r' \in [0, 1]$ and (12) must be satisfied, the terms $p'-q'$, $q'-q$, and $q-p$ cannot 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:

$$\begin{array}{llllll}
(1) & \begin{cases} p'-q' > 0, \\ q-p > 0, \\ q'-q < 0, \end{cases} & (2) & \begin{cases} p'-q' > 0, \\ q-p < 0, \\ q'-q > 0, \end{cases} & (3) & \begin{cases} p'-q' < 0, \\ q-p > 0, \\ q'-q > 0, \end{cases} \\
(4) & \begin{cases} p'-q' < 0, \\ q-p < 0, \\ q'-q > 0, \end{cases} & (5) & \begin{cases} p'-q' < 0, \\ q-p > 0, \\ q'-q < 0, \end{cases} & (6) & \begin{cases} p'-q' > 0, \\ q-p < 0, \\ q'-q < 0. \end{cases}
\end{array} \qquad (26)$$

As an illustration, we present in the next section all the details of the analysis of the Gini Index and Information Gain functions for one of the six cases enumerated in (26) and we give a summary of the remaining five cases.

5. Case study

In this section we will use the previously introduced methodology to compare the behavior of the Gini Index and Information Gain criteria in the first case defined in (26). We present the intervals of coincidence/non-coincidence in the choice of the split attribute for the two 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 first possible situation. The complete analysis can be found in [17]. If the sign of the difference of the Gini Index functions $gini(T) - gini(T')$ in (18) is the same as the sign of the difference of the Information Gain functions $gain(T) - gain(T')$ in (23), then the two split criteria select the same attribute to split on, otherwise they select different attributes to split on.

Case 1. $p'-q' > 0, q-p > 0, q'-q < 0.$

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). $0 < p < q' < q < p' < 1$.

Proof. We have to establish in this subcase the sign of the expression (18). First we show that $f_1 \in (0, 1)$. We have $f_1 > 0$ as $q' - q < 0$, $q - p' < 0$, $p - q < 0$ and $p + q - p' - q' < 0$. Using the expression of f_1 given in (19) we obtain: $f_1 - 1 = \frac{(q'-p)(p-p')}{(p-q)(p+q-p'-q')} < 0$ as $q' - p > 0$, $p - p' < 0$, $p - q < 0$, and $p + q - p' - q' < 0$.

For r and r' we must assure that conditions (13)–(16) are satisfied. (15) is satisfied as: $p - q < 0$, $q - p' < 0$, $p' - q' > 0$ and $r \geqslant 0$. (16) is satisfied as: $p' - q' > 0$, $q' - p > 0$, $p - q < 0$ and $r' \geqslant 0$. But to verify that (13) and (14) are satisfied, it is necessary that $r \leqslant \frac{q'-q}{p-q}$ and $r' \leqslant \frac{q-q'}{p'-q'}$. 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: $r \in [0, \frac{q'-q}{p-q}]$ and $r' \in [0, \frac{q-q'}{p'-q'}]$.

In addition we can easily show that $f_1 < \frac{q'-q}{p-q}$ as $f_1 - \frac{q'-q}{p-q} = \frac{q-q'}{p+q-p'-q'}$ and we have that $q - q' > 0$ and $p + q - p' - q' < 0$.

Knowing 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 $gini(T')$ given by (18). For $r \in [0, f_1]$ we have $gini(T) - gini(T') \leqslant 0$ and for $r \in [f_1, \frac{q'-q}{p-q}]$ we have $gini(T) - gini(T') \geqslant 0$.

To evaluate the difference between $gain(T)$ and $gain(T')$ expressed in (23) 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 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 (25) to the probabilities $p < q < p'$, $p < q' < q$, $p < q' < p'$, 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 would contradict the monotonicity of f'.

Now, it is easy to show that $E_1 \geqslant 0$, $E_2 \leqslant 0$ and $f_2 \in [0, \frac{q'-q}{p-q})$. We have $f_2 < \frac{q'-q}{p-q}$ as using (24):

$$f_2 < \frac{q'-q}{p-q} \quad\Longleftrightarrow\quad \frac{f'(x_2) - f'(x_3)}{f'(x_2) - f'(x_1)} < 1 \quad\Longleftrightarrow\quad f'(x_3) > f'(x_1)$$

which is true as demonstrated before. So, for $r \in [0, f_2]$ we have $gain(T) - gain(T') \leqslant 0$, and for $r \in [f_2, \frac{q'-q}{p-q}]$ we have $gain(T) - gain(T') \geqslant 0$.

In conclusion, for $0 < p < q' < q < p' < 1$ we have: $r \in [0, \frac{q'-q}{p-q}]$, $r' \in [0, \frac{q-q'}{p'-q'}]$ and $f_1, f_2 \in [0, \frac{q'-q}{p-q}]$. If $r \in [0, \min\{f_1, f_2\}]$, then the same test T' is selected by both split criteria. If $r \in (\min\{f_1, f_2\}, \max\{f_1, f_2\})$, then different splits are selected. If $r \in [\max\{f_1, f_2\}, \frac{q'-q}{p-q}]$, then the same test T is selected by both split criteria. □

Case 1(b). $0 < p < q' < p' < q < 1$.

Proof. To establish the position of r and r' we use conditions (13)–(16) as we did before, and we obtain: $r \in [\frac{p'-q}{p-q}, \frac{q'-q}{p-q}]$ and $r' \in [0, 1]$. The expression $p + q - p' - q'$ can be negative or positive. If $p+q-p'-q' \leqslant 0$ then, as $r \geqslant 0$, $p-q < 0$, $q-q' > 0$, and, $q - p' > 0$, we have $gini(T) - gini(T') \geqslant 0$. If $p + q - p' - q' \geqslant 0$ then we have $f_1 \geqslant 1$. As $r \leqslant 1$, then we have $r \leqslant f_1$ and, therefore we have $gini(T) - gini(T') \geqslant 0$. Therefore, for $r \in [\frac{p'-q}{p-q}, \frac{q'-q}{p-q}]$ and $r' \in [0, 1]$ we have $gini(T) - gini(T') \geqslant 0$.

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 1(a). Applying the proposition (25) to the probabilities $p < q' < p'$, $p < q' < q$, $q' < p' < q$, 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 $x_2 < x_1 < x_3$.

(i) In the case $x_1 < x_2 < x_3$, we have that $f'(x_1) < f'(x_2) < f'(x_3)$, therefore $E_1 \geqslant 0$ and $E_2 \geqslant 0$. Thus we have $gain(T) - gain(T') \geqslant 0$.

(ii) In the case $x_2 < x_1 < x_3$, we have that $f'(x_2) < f'(x_1) < f'(x_3)$, so $E_1 \leqslant 0$, $E_2 \geqslant 0$. We show that

$$f_2 > \frac{q'-q}{p-q}: \quad f_2 > \frac{q'-q}{p-q} \iff \frac{f'(x_2) - f'(x_3)}{f'(x_2) - f'(x_1)} > 1$$
$$\iff f'(x_3) > f'(x_1)$$

which is true as we are in the case $x_2 < x_1 < x_3$ and f' is strictly monotonically increasing. Therefore we have $gain(T) - gain(T') \geqslant 0$.

In conclusion, for $0 < p < q' < p' < q < 1$ we have: $r \in [\frac{p'-q}{p-q}, \frac{q'-q}{p-q}]$, $r' \in [0, 1]$, 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 $r' \in [\frac{p-q'}{p'-q'}, \frac{q-q'}{p'-q'}]$. The expression $p + q - p' - q'$ can be negative or positive. If $p + q - p' - q' \geqslant 0$ then, as $r \geqslant 0$, $p - q < 0$, $q - q' > 0$, and $q - p' < 0$, we have $gini(T) - gini(T') \leqslant 0$. If $p + q - p' - q' \leqslant 0$ then we have $f_1 \geqslant 1$. As $r \leqslant 1$, then we have $r \leqslant f_1$ and, therefore we have $gini(T) - gini(T') \leqslant 0$. For $r \in [0, 1]$ and $r' \in [\frac{p-q'}{p'-q'}, \frac{q-q'}{p'-q'}]$ we have $gini(T) - gini(T') \leqslant 0$.

Applying proposition (25) to the probabilities $q' < p < q$, $q' < p < p'$, $q' < q < p'$, 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 $x_3 < x_2 < x_1$.

(i) If $x_3 < x_1 < x_2$ we have: $E_1 \geqslant 0$, $E_2 \leqslant 0$, and $f_2 > 1$, and therefore, $gain(T) - gain(T') < 0$. To demonstrate that $f_2 > 1$ is equivalent to show that

$$\frac{f'(x_2) - f'(x_3)}{f'(x_2) - f'(x_1)} > \frac{q - p}{q - q'}.$$

The left-hand side of the inequality is greater than 1 as we have

$$\frac{f'(x_2) - f'(x_3)}{f'(x_2) - f'(x_1)} > 1 \quad \Longleftrightarrow \quad f'(x_3) < f'(x_1).$$

The right-hand side of the inequality is strictly less than 1 as we have

$$\frac{q - p}{q - q'} < 1 \quad \Longleftrightarrow \quad p > q'.$$

By combining these two observations the inequality to show becomes obviously.

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

In conclusion, for $0 < q' < p < q < p' < 1$ we have: $r \in [0, 1]$, $r' \in [\frac{p-q'}{p'-q'}, \frac{q-q'}{p'-q'}]$, 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. Using the conditions (13)–(16) we obtain: $r \in [\frac{p'-q}{p-q}, 1]$, $r' \in [\frac{p-q'}{p'-q'}, 1]$ and $f_1 \in [\frac{p'-q}{p-q}, 1]$. If $r \in [\frac{p'-q}{p-q}, f_1]$, then we have $gini(T) - gini(T') \geqslant 0$. If $r \in [f_1, 1]$, then we have $gini(T) - gini(T') \leqslant 0$.

Applying the proposition (25) to the probabilities $q' < p < p'$, $q' < p < q$, $q' < p' < q$, 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: $E_1 \leqslant 0$, $E_2 \geqslant 0$, and $f_2 \in (\frac{p'-q}{p-q}, 1)$. For $r \in [\frac{p'-q}{p-q}, f_2]$ we have $gain(T) - gain(T') \geqslant 0$ and for $r \in [f_2, 1]$ we have $gain(T) - gain(T') \leqslant 0$.

(i) *Proof for* $f_2 > \frac{p'-q}{p-q}$. We demonstrate this inequality by equivalencies. First we use the expression of f_2 given in (24) and we obtain:

$$f_2 > \frac{p' - q}{p - q} \quad \Longleftrightarrow \quad \bigl(f'(x_3) - f'(x_2)\bigr)\bigl(q - q'\bigr) > \bigl(f'(x_1) - f'(x_2)\bigr)\bigl(q - p'\bigr).$$

After some simple calculations we obtain:

$$\Longleftrightarrow \quad f'(x_3)\bigl(q - q'\bigr) + f'(x_2)\bigl(q' - p'\bigr) + f'(x_1)\bigl(p' - q\bigr) > 0.$$

We substitute $f'(x_1)$, $f'(x_2)$, $f'(x_3)$ by using (20)–(22):

$$\begin{aligned}
\iff\quad & \frac{f(q)-f(q')}{q-q'}\big(q-q'\big)+\frac{f(q')-f(p')}{q'-p'}\big(q'-p'\big)\\
& +\frac{f(q)-f(p)}{q-p}\big(p'-q\big)>0\\
\iff\quad & f(q)\big(p'-p\big)-f(p)\big(p'-q\big)-f\big(p'\big)(q-p)>0. \qquad (*)
\end{aligned}$$

As $p < p' < q$, $\exists\alpha \in (0, 1)$, so that

$$\begin{aligned}
& p' = \alpha p + (1-\alpha)q,\\
& (*) \quad\iff\quad f(q)(1-\alpha)+f(p)\alpha > f\big(\alpha p+(1-\alpha)q\big)
\end{aligned}$$

which is true by the strict convexity of f.

(ii) *Proof for* $f_2 < 1$. As we did before, we use also here the proof by equivalencies. We substitute f_2 by its expression given in (24) and we obtain:

$$\begin{aligned}
f_2 < 1 \quad\iff\quad & \big(f'(x_2)-f'(x_3)\big)\big(q-q'\big) > \big(f'(x_2)-f'(x_1)\big)(q-p)\\
\iff\quad & f'(x_3)\big(q'-q\big)+f'(x_2)\big(p-q'\big)+f'(x_1)(q-p)>0.
\end{aligned}$$

We substitute $f'(x_1)$, $f'(x_2)$, $f'(x_3)$ by their expressions given in (20)–(22):

$$\begin{aligned}
\iff\quad & \frac{f(q')-f(q)}{q'-q}\big(q'-q\big)+\frac{f(p')-f(q')}{p'-q'}\big(p-q'\big)\\
& +\frac{f(q)-f(p)}{q-p}(q-p)>0\\
\iff\quad & f\big(q'\big)\big(p'-p\big)-f(p)\big(p'-q'\big)+f\big(p'\big)\big(p-q'\big)>0. \qquad (**)
\end{aligned}$$

As $q' < p < p'$, $\exists\, \alpha \in (0, 1)$, so that:

$$\begin{aligned}
& p = \alpha q' + (1-\alpha)p'\\
& (**) \quad\iff\quad f\big(p'\big)(1-\alpha)+f\big(q'\big)\alpha > f\big(\alpha q'+(1-\alpha)p'\big)
\end{aligned}$$

which is true by the strict convexity of f.

In conclusion, for $0 < q' < p < p' < q < 1$ we have $r \in [\frac{p'-q}{p-q}, 1]$ and $r' \in [\frac{p-q'}{p'-q'}, 1]$. For $r \in [\frac{p'-q}{p-q}, \min\{f_1, f_2\}]$ the same test T is selected by both split criteria, for $r \in (\min\{f_1, f_2\}, \max\{f_1, f_2\})$ different splits are selected, and for $r \in [\max\{f_1, f_2\}, 1]$ the same test T' is selected by both split criteria. □

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

Proof. This case is dropped as it contradicts to conditions (15) and (16). As $p' - q' < 0$,

$$(15) \iff r(p-q) + q - p' \leqslant 0 \iff r \geqslant \frac{p'-q}{p-q}.$$

As the ratio $\frac{p'-q}{p-q}$ is strictly greater than 1, this implies that $r > 1$, but this is in contradiction with the fact that r represents a probability, so that r must be equal or less than 1. Therefore such a case cannot be possible. □

We do not present here the analysis of the remaining five cases enumerated in (26). 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 [17].

Here we present a synthesis of the obtained results. Suppose we have two available tests T, T' and our task is to determine if the test selected by the Gini Index or Information Gain criterion is the same or not. T and T' can be characterized by the parameters p, q, r and p', q', r', respectively. We determine the maximum and the minimum of the probabilities $\{p, q, p', q'\}$:

(i) If $\max\{p, q, p', q'\}$ and $\min\{p, q, p', q'\}$ belong to the same test, i.e., we obtain $\{p, q\}$, or $\{p', q'\}$ as the pair of minimum and maximum, then the two criteria of split will select the same test to split on.

(ii) If we obtain $\{p, p'\}$, $\{p, q'\}$, $\{q, p'\}$ or $\{q, q'\}$ as pair of minimum and maximum, then there are two possible situations to analyze. If $(f_1 - r)(f_2 - r) > 0$, then the two criteria choose the same test, and, if $(f_1 - r)(f_2 - r) < 0$, then the two criteria choose different tests.

The results obtained for the six cases identified can be summarized in the following way. For the case (1) we obtained two situations in which the two split criteria select 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 one 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 they are choosing different tests. 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 considered all possible databases having two binary attributes and one binary decision attribute containing 50–200 tuples. We calculated then for all sizes of databases the number of cases of disagreement. The number of cases of disagreement was never higher than 2% of all cases. This explains why most empirical studies concluded that there is no significant difference between the two criteria. Of course this does not mean that for some specific databases there might be significant differences.

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/disagreement of the Gini Index function and the Information Gain function. We found that they disagree only in 2% of all cases, which explains why most previously published empirical results concluded that it is not possible to decide which one of the two tests performs better. 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 [20].

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) (1992) 373–387.

[2] E. Baker and A.K. Jain, On feature ordering in practice and some finite sample effects, in: *Proceedings of the Third International Joint Conference on Pattern Recognition*, San Diego, CA (1976) pp. 45–49.

[3] M. Ben-Bassat, Myopic policies in sequential classification, IEEE Transactions on Computing 27(2) (1978) 170–174.

[4] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees* (Wadsworth International Group, 1984).

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

[6] J. Gama and P. Brazdil, Characterization of classification algorithms, in: *EPIA-95: Progress in Artificial Intelligence, 7th Portuguese Conference on Artificial Intelligence*, eds. C. Pinto-Ferreira and N. Mamede (Springer, 1995) pp. 189–200.

[7] I. Kononenko, On biases in estimating multi-valued attributes, in: *IJCAI-95: Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, ed. C. Mellish (Morgan Kaufmann, San Mateo, CA, 1995) pp. 1034–1040.

[8] T.-S. Lim, W.-Y. Loh and Y.-S. Shih, A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms, Machine Learning (1999).

[9] J. Mingers, Expert systems-rule induction with statistical data, Journal of the Operational Research Society 38(1) (1987) 39–47.

[10] J. Mingers, An empirical comparison of selection measures for decision tree induction, Machine Learning 3 (1989) 319–342.

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

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

[13] K.V.S. Murthy, On growing better decision trees from data, Ph.D. thesis, The John Hopkins University, Baltimore, MD (1995).

[14] G. Pagallo, Adaptive decision tree algorithms for learning from examples, Ph.D. thesis, University of California, Santa Cruz, CA (1990).

[15] J.R. Quinlan, Simplifying decision trees, International Journal of Man–Machine Studies 27 (1987) 221–234.

[16] J.R. Quinlan, *C4.5 Programs for Machine Learning* (Morgan Kaufmann, 1993).

[17] L.E. Raileanu, Formalization and comparison of split criteria for decision trees, Ph.D. thesis, University of Neuchâtel, Switzerland (May 2002).

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

[19] M. Sahami, Learning non-linearly separable Boolean functions with linear threshold unit trees and madaline-style networks, in: *Proceedings of the Eleventh National Conference on Artificial Intelligence* (AAAI Press, 1993) pp. 335–341.

[20] K. Stoffel and L.E. Raileanu, Selecting optimal split-functions for large datasets, in: *Research and Development in Intelligent Systems XVII*, BCS Conference Series (2000).

[21] R. Vilalta and D. Oblinger, A quantification of distance-bias between evaluation metrics in classification, in: *Proceedings of the 17th International Conference on Machine Learning*, Stanford University (2000).

[22] A.P. White and W.Z. Liu, Bias il information-based measures in decision tree induction, Machine Learning 15(3) (1997) 321–328.