To be presented at IPMU'96
Info. Proc. and Manag. of Uncertainty in Knowledge-Based Systems
July 1-5, 1996, Granada, Spain

On Uncertainty Measures Used for Decision Tree Induction

Louis Wehenkel
Research Associate of the National Fund for Scientific Research
Department of Electrical Engineering - University of Liège
Institut Montefiore - Sart-Tilman B28 - B-4000 Liège - Belgium
lwh@montefiore.ulg.ac.be

Abstract

This paper provides a further look at uncertainty or information criteria used in the context of decision tree induction, and more generally of learning conditional class probability models. We show the high degree of similarity among two main families of criteria based respectively on the logarithmic *SHANNON* entropy function and the quadratic *GINI* index. We start by introducing a general family of entropy functions and then discuss the latter particular cases, and end up with a short review of the Kolmogorov-Smirnov distance, another related measure.

1 Generalized Information Functions

The concept of generalized information functions of type β was first introduced by Daróczy [1] and its use for pattern recognition problems was discussed by Devijver [2].

The entropy of type β (β positive and different from 1) of a discrete probability distribution $(p_1, \ldots, p_m)$ is defined by

$$H^\beta(p_1, \ldots, p_m) \triangleq \sum_{i=1}^{m} p_i u^\beta(p_i), \qquad (1)$$

where the uncertainty $u^\beta(p_i)$ is defined by

$$u^\beta(p_i) \triangleq \frac{2^{\beta-1}}{2^{\beta-1}-1}(1-p_i^{\beta-1}). \qquad (2)$$

Measure u^β is a strictly decreasing function of p_i.

1.1 Fundamental Properties of H^β [1, 2, 3]

$\boxed{\text{P1}}$ $H^\beta(p_1, \ldots, p_m) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left[1 - \sum_{i=1}^{m} p_i^\beta\right]$;

$\boxed{\text{P2}}$ $H^\beta(p_1, \ldots, p_m)$ is invariant with respect to the permutation of its arguments;

$\boxed{\text{P4}}$ $H^\beta(1) = H^\beta(0, \ldots, 1, \ldots, 0) = 0$ and $H^\beta(\frac{1}{2}, \frac{1}{2}) = 1$;

$\boxed{\text{P5}}$ $H^\beta(p_1, \ldots, p_{m-1}, p_m) = H^\beta(p_1, \ldots, p_{m-1} + p_m) + (p_{m-1} + p_m)^\beta H^\beta(p_{m-1}/(p_{m-1} + p_m), p_m/(p_{m-1} + p_m))$ (pseudo-additivity);

$\boxed{\text{P6}}$ $0 \leq H^\beta(p_1, \ldots, p_m) \leq H^\beta(\frac{1}{m}, \ldots, \frac{1}{m})$, i.e. maximum entropy corresponds to the uniform distribution;

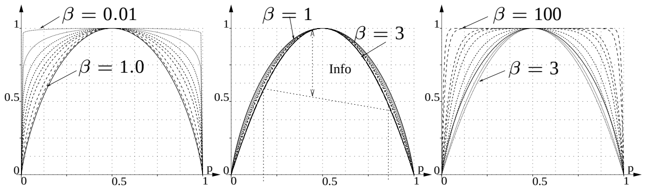

Figure 1: Entropy functions H^β for $\beta \in [0.01 \ldots 100.0]$

$\boxed{\text{P7}}$ $H^\beta(p_1 \ldots p_m)$ is a concave ($\cap$) function[1] :

$$\forall \lambda_j \geq 0, p_{ij} \geq 0 \ : \sum_{j=1}^{k} \lambda_j = 1 \wedge \sum_{i=1}^{m} p_{ij} = 1 \ \Rightarrow$$

$$H^\beta(\sum_{j=1}^{k} \lambda_j p_{1j}, \ldots, \sum_{j=1}^{k} \lambda_j p_{mj}) \geq \sum_{j=1}^{k} \lambda_j H^\beta(p_{1j}, \ldots, p_{mj}).$$

The first five properties follow directly from the definition of H^β. The proofs of the other two properties are not given here to save space. The *convexity* property (P7) has many important implications further commented below.

Daróczy shows that properties 2, 4 and 5 provide a complete characterization of the entropy functions of type β. Figure 1 reproduces H^β in the two-class case ($p_1 = p; p_2 = 1-p$), for a large range of β values. In particular, the quadratic entropy ($\beta = 2$) and the logarithmic one ($\beta \to 1$) are hardly distinguishable.

Note that *additivity* of entropies, i.e.

$$H^\beta(p_1, \ldots, p_{m-1}, p_m) = H^\beta(p_1, \ldots, p_{m-1} + p_m)$$
$$+ (p_{m-1} + p_m) H^\beta\left(\frac{p_{m-1}}{(p_{m-1} + p_m)}, \frac{p_m}{(p_{m-1} + p_m)}\right),$$

can only achieved by letting β converge towards 1, yielding the logarithmic entropy (see §2).

1.2 Conditional entropies

Let t and c denote two discrete random variables (e.g. t might denote a test issue at a node of a decision tree, and c the class which we would like to predict by the tree) of respective probability distribution $(p(t_1), \ldots, p(t_k))$ and $(p(c_1), \ldots, p(c_m))$. We denote by

$$H_C^\beta \triangleq H^\beta(p(c_1), \ldots, p(c_m)), \qquad (3)$$

[1] on the convex set defined by $p_i \geq 0$ and $\sum_{i=1}^{m} p_i = 1$

the prior classification entropy of type β and the conditional type β entropy is defined by

$$H^{\beta}_{C|t_j} \triangleq H^{\beta}_C(p(c_1 \mid t_j), \ldots, p(c_m \mid t_j)), \qquad (4)$$

and the mean conditional type β entropy by

$$H^{\beta}_{C|T} \triangleq \sum_{j=1}^{k} p(t_j) H^{\beta}_{C|t_j}. \qquad (5)$$

The concave nature of H^{β} implies the following fundamental monotonicity property (see e.g. [3] for a proof)

$$H^{\beta}_{C|T} \leq H^{\beta}_C. \qquad (6)$$

Furthermore, due to the strictness of the concavity the following equality holds also

$$H^{\beta}_{C|T} = H^{\beta}_C \;\Leftrightarrow\; p(c_i|t_j) = p(c_i), \;\forall\; i, j; \qquad (7)$$

i.e. if and only if c and t are statistically *independent*.

2 Shannon Entropy

For $\beta = 1$, $u^{\beta}(x)$ is not defined. However, since $\lim_{\beta \to 1} u^{\beta}(x) = -\log_2 x$

$$H \triangleq \lim_{\beta \to 1} H^{\beta} = -\sum_{i=1}^{m} p_i \log_2 p_i, \qquad (8)$$

i.e. the well-known logarithmic or Shannon entropy

It may be easily checked that properties (P1-P7) still hold for the logarithmic entropy. In particular (P5) now expresses additivity, i.e. the fact that the entropy of two independent events is equal to the sum of their respective entropies.

The logarithmic entropy is the basis for various interpretations in the context of probabilistic modeling (likelihood of the data given a model / posterior probability of a model given the data and model priors) [3, 4, 5]. Let us merely note that these interpretations are certainly among the main reasons of the high popularity of this particular uncertainty measure [6].

2.1 Conditional entropies and information

The mean conditional entropy becomes

$$H_{C|T} = -\sum_{j=1}^{k} \sum_{i=1}^{m} p(c_i, t_j) \log_2 p(c_i \mid t_j). \qquad (9)$$

The following quantities of interest are also defined.

- The entropy of t,

$$H_T = -\sum_{j=1}^{k} p(t_j) \log_2 p(t_j) \qquad (10)$$

- The mean conditional entropy of t given c

$$H_{T|C} = -\sum_{i=1}^{m} \sum_{j=1}^{k} p(c_i, t_j) \log_2 p(t_j \mid c_i). \qquad (11)$$

- The joint entropy of t and c

$$H_{C,T} = -\sum_{i=1}^{m} \sum_{j=1}^{k} p(c_i, t_j) \log_2 p(c_i, t_j). \qquad (12)$$

- The mutual informations

$$\begin{aligned} I^T_C &\triangleq H_C - H_{C|T}, & (13)\\ &= -\sum_{i=1}^{m} \sum_{j=1}^{k} p(c_i, t_j) \log_2 \frac{p(c_i)}{p(c_i|t_j)}, & (14)\\ I^C_T &\triangleq H_T - H_{T|C}, & (15)\\ &= -\sum_{i=1}^{m} \sum_{j=1}^{k} p(c_i, t_j) \log_2 \frac{p(t_j)}{p(t_j|c_i)}. & (16) \end{aligned}$$

The following relationships are satisfied.

- Additivity of entropies

$$H_{C,T} = H_C + H_{T|C} = H_T + H_{C|T} = H_{T,C}. \qquad (17)$$

- And consequently reciprocity of mutual information

$$\begin{aligned} I^T_C &= H_C - H_{C|T} \;=\; H_T + H_C - H_{T,C} \\ &= H_T - H_{T|C} \;=\; I^C_T. \qquad (18) \end{aligned}$$

- Thus,

$$I^C_T = -\sum_{i=1}^{m} \sum_{j=1}^{k} p(c_i, t_j) \log_2 \frac{p(c_i)p(t_j)}{p(c_i, t_j)}. \qquad (19)$$

- Inequalities

$$\begin{aligned} &H_{T|C} \leq H_T; H_{C|T} \leq H_C \;;\; I^T_C \leq H_C; \\ &I^T_C \leq H_T \;;\; I^T_C \leq H_{C,T} \;;\; I^T_C \geq 0. \qquad (20) \end{aligned}$$

Further, under the necessary and sufficient condition of strict association between t and c (i.e. $p(c_i, t_j)$ diagonalized by permutation of columns or lines) the following equalities hold

$$I^T_C = H_T = H_C = H_{C,T}; H_{T|C} = H_{C|T} = 0. \qquad (21)$$

Finally, under the necessary and sufficient condition of statistical independence the following equalities hold.

$$H_T = H_{T|C}; H_C = H_{C|T}; H_{C,T} = H_C + H_T; I^T_C = 0. \qquad (22)$$

2.2 Normalized information

I^T_C measures the reduction of the uncertainty of one of the variables t or c, given the knowledge of the other one. In the context of decision tree induction it is useful as an evaluation function of alternative tests at a tree node, in order to select the one reducing most significantly the uncertainty about the unknown classification. More generally, in the context of statistical modeling this measure may be used to assess the information provided by alternative models.

Within this context, the fact that the information quantity is upper bounded by the prior entropy H_C makes its interpretation difficult. The prior entropy, and hence the information of candidate models will indeed be highly variable according to the number and prior probabilities of classes.

Another frequently mentioned difficulty in the context of decision tree induction concerns the bias of estimates of information which increases with k and m. This tends to favor tests at a tree node with a larger number of outcomes [3, 7, 8] (i.e. higher k).

Thus, various normalized “correlation” measures have been derived from the information quantity so as to yield improved “score” measures [7, 8, 9, 10]. We will discuss some of them below and provide an illustration on the basis of data related to a practical example.

2.2.1 Normalization by H_C

We denote this score measure by
$$A_C^T \triangleq \frac{I_C^T}{H_C}. \tag{23}$$

In the context of decision tree building, at a given tree node H_C is constant. Thus the ranking provided by A_C^T and I_C^T are equivalent and the normalization has no effect at all on the resulting tree. We use it here in place of I_C^T, merely for comparison purposes.

It is worth mentioning that I_C^T and consequently A_C^T presents at least two interesting properties which do not hold necessarily for the other measures presented below.

The first property concerns the location of optimal thresholds for ordered attributes. One may indeed show that for ordered attributes, the optimal thresholds maximizing I_C^T must lie at so-called cut-points, i.e. values where the class probabilities are not stationary [11]. (In the finite sample case, this excludes in particular all thresholds lying between objects of identical classes.) Exploiting this property allows in general to reduce significantly the computational burden of searching for the optimal thresholds.

The second property concerns the search for an optimal binary partition for a qualitative attribute [12, 13]. It allows to reduce the search from $2^{L-1}-1$ to L candidate partitions (where L denotes the number of different values assumed by the qualitative attribute).

2.2.2 Normalization by H_T

In order to reduce the bias towards many-valued splits, Quinlan introduced the so-called "gain ratio" [8], which we denote by
$$B_C^T \triangleq \frac{I_C^T}{H_T}. \tag{24}$$

Dividing by H_T allows to reduce the bias of I_C^T towards tests with many successors (yielding a high value of H_T).

However, a possible problem with this measure lies in the fact that it may overestimate the value of splits with very low H_T values, in particular splits corresponding to uneven decompositions of a learning set into subsets. Thus, for ordered attributes the optimal values of B_C^T often tend to be located closer to its extreme values; this is known in the literature as the "end-cut" preference of the "gain ratio" criterion (see also the example in §2.2.5 below).

2.2.3 Normalization by $\frac{1}{2}(H_C + H_T)$

The preceding normalizations yield asymmetrical "score" measures. While it has been suggested that asymmetrical measures are natural in the context of pattern recognition applications, because the learning objective privileges the classification variable [2], we believe that symmetrical measures are more appropriate. Indeed, in the context of decision tree building a main objective is assessment of correlations among attributes and classifications, and also among various attributes. There is no reason that the correlation of two attributes should not be symmetrical.

Thus, sharing the opinion of Kvålseth [9], we advocate the use of the following measure [14].
$$C_C^T \triangleq \frac{2I_C^T}{H_C + H_T}, \tag{25}$$
which is symmetrical in C and T.

Kvålseth shows that if $I_C^T > 0$, the sampling estimate $\hat{C}_C^T$ is asymptotically normally distributed with mean C_C^T and thus is (asymptotically) unbiased. One of its main practical advantages is that Kvålseth provides the following explicit formulation of its standard deviation
$$\sigma_{C_C^T} = \sqrt{\left(\frac{C_C^t}{n_{..}I_C^t}\right)^2 \sum_{i=1,m}\sum_{j=1,p} n_{ij}\Bigg[\log n_{ij} + \left(\frac{C_C^t}{2}-1\right) \log(n_{i.}n_{.j}) + (1-C_C^t)\log n_{..}\Bigg]^2}, \tag{26}$$
where $n_{..}$ denotes the sample size, n_{ij} the expected number of samples of class c_j yielding test issue t_j (i.e. $n_{ij} = n_{..}p(c_i,t_j)$), $n_{i.} \triangleq \sum_j n_{ij}$ and $n_{.j} \triangleq \sum_i n_{ij}$.)

Equation. (26) evaluates the level of inaccuracy of the sample estimate of the uncertainty measure. This provides valuable information in order to assess the significance of score differences among various candidate models. The sample estimate of $\sigma_{C_C^T}$ is obtained by replacing in eqn. (26) the expected numbers $n_{i,j}$ by their sample estimates (i.e. by the cell counts).

2.2.4 Normalization by $H_{C,T}$

Another symmetrical and normalized measure more recently proposed by López de Mántaras is defined by [7]
$$D_C^T \triangleq \frac{I_C^T}{H_{C,T}}. \tag{27}$$

This author shows formally that D_C^T is not biased towards many-valued splits, and suggests also that it tends to provide simpler trees than the gain ratio measure. He shows also that $1 - D_C^T$ is a proper distance measure of two probability distributions $(p(c_1),\ldots,p(c_m))$ and $(p(t_1),\ldots,p(t_k))$, which satisfies the triangular inequality.

Let us show the equivalence of the last two measures C_C^T and D_C^T.

Noting that $H_{C,T} = H_C + H_T - I_C^T$ we find that
$$D_C^T = \frac{I_C^T}{H_C + H_T - I_C^T}, \tag{28}$$
or equivalently that
$$D_C^T = \frac{1}{\frac{H_C+H_T}{I_C^T} - 1} \tag{29}$$
Thus
$$D_C^T = \frac{1}{\frac{2}{C_C^T} - 1}, \tag{30}$$
which implies that the two measures are a monotonic transformation of each other, as shown in Fig. 2. Hence the preference relationship induced by these measures are identical. Therefore, as far as the *ranking* of candidate tests is concerned the formal property of no bias towards multiple-valued splits of D_C^T must also hold for C_C^T.

2.2.5 Comparison

First of all we recall that it has been reported many times from experimental studies that the predictive classification

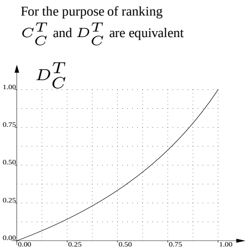

Figure 2: D_C^T vs C_C^T

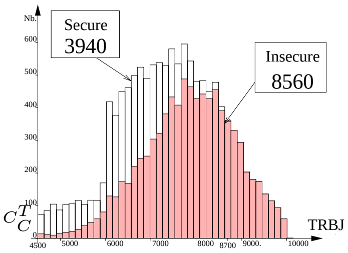

Figure 3: Trbj vs security

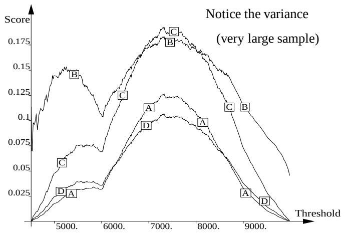

Figure 4: Scores vs Threshold

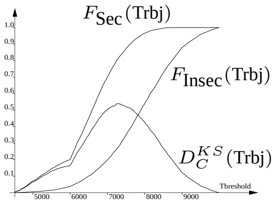

Figure 5: D_C^{KS} vs Threshold

reliability of decision trees appears to be not much affected by the type of attribute selection criteria used [7, 10, 12, 15].

However, the complexity of the trees and hence their interpretability which is one of their main attractive features, does often depend much more strongly on the type of measure used. Further, since the complexity of the tree will influence the size of the learning samples at its terminal nodes, it will influence strongly the accuracy of their class probability estimates. Information about accuracy of class probability estimates and tree complexities is however seldom reported in experimental studies. In addition the value of simplicity may depend on pragmatic considerations which are difficult to take into account in blind comparisons.

Let us consider an example data base of a real life electric power transient stability assessment problem, taken from [16]. It is composed of 12,500 randomly generated operating conditions of the EHV power system of Hydro-Québec, classified into secure and insecure classes by numerical simulation. A state is considered as insecure if there exists a plausible disturbance which would lead to a loss of transient stability. In addition, every state is described by about 100 attributes which provide information about the electrical and topological situation : power flows, number of lines in operation in different corridors, generation in different power plants and automatic voltage regulating devices in operation. The objective of applying a decision tree induction algorithm to this problem is to identify among these parameters those having a stronger influence on the security and to formulate automatically operating strategies expressed in terms of these variables.

For the purpose of illustration Fig. 3 depicts the frequency distribution of one power flow attribute (denoted by Trbj) which is found to strongly influence the security of the system. The horizontal axis represents the value of the attribute, and the overall height of the bars is proportional to the number of states among the 12,500 which lie in the corresponding range of attribute values; the relative height of the dark and the light regions is proportional to the relative frequency of insecure and secure states (estimating the conditional class probabilities : $p(c_i|\text{Trbj} \in [x, x + \Delta\text{Trbj}[$.)

Let us consider the case where we use a test on this attribute in order to discriminate among secure and insecure states. In other words, for a given threshold, we define t as a binary random variable, where $t = t_1$ if Trbj $<$ Threshold, and $t = t_2$ otherwise. Varying the threshold between its minimum and maximum values then defines a family of tests, and for a given score measure, the optimal threshold will be the one leading to a maximum score, as estimated on the basis of the sample of 12,500 states. Fig. 4 represents the variation of the four measures (A, B, C, and D) as a function of the test threshold. In spite of the very large sample size, it is possible to observe the small non-smooth random fluctuations.

From the observation of these curves we draw the following comments. First of all, all four measures present two salient local maxima, one below 6000MW and one around 7300MW, which is also the global maximum. Actually, they correspond to the two different statistical populations from which the data base samples where drawn. In addition to these dominant tendencies, there are small high frequency oscillations translating the effect of the sampling of the probability distributions of classes. They vanish however above 8700MW, where all four curves start decreasing monotonically. This is merely the consequence of the fact that above this threshold value all the states of the data base belong to the same class (see Fig. 3 and Fig. 5).

Comparing the curve related to measure A with the three others, we observe that the normalization of B, C and D taking into account H_T, enhances indeed the scores nearby the upper and lower bound of the threshold interval. In particular, the value of the local maximum nearby 5700MW is enhanced, and pulled towards the smaller threshold values. This effect is stronger for measure B_C^T than for measures C_C^T and D_C^T. Incidentally, we note that the latter two measures are indeed equivalent, in terms of the location of all the local maxima of their curve.

Finally, we may observe in this present example the odd behavior of measure B_C^T near the extreme values of the threshold interval, where $H_T \approx 0$. In particular its limit value is not equal to zero. As a conclusion we would not recommend this type of normalization.

2.3 Hypothesis testing

Here we merely recall the well known fact that under the hypothesis of statistical independence the finite sample estimate $2n_{..} \ln 2\hat{I}_C^T$ is distributed according to a $\chi-$square law of $(m-1) \times (k-1)$ degrees of freedom [9].

Thus $\hat{I}_C^T$ will assume the following expected value

$$E\{\hat{I}_C^T\} = \frac{(m-1)(k-1)}{2n_{..} \ln 2}. \tag{31}$$

This confirms[2] the fact that I_C^T is biased, and the higher the number of successors and classes, the higher the bias. On

[2] strictly speaking only under the independence hypothesis

the other hand, the bias decreases towards zero when the sample size $n_{..}$ increases.

3 Quadratic Entropy

Setting $\beta = 2$ in eqn. (1) and (2) yields the quadratic entropy

$$H^2 = 2\left[1 - \sum_{i=1}^{m} p_i^2\right] \quad (32)$$

$$H^2 = 4\sum_{i \neq j} p_i p_j = 2\sum_{i=1}^{m} p_i(1-p_i). \quad (33)$$

This is identical to the so-called "Gini" index [12], which may be interpreted in the following way. Let us suppose that an object is classified randomly into c_i, with a probability equal to $p(c_i)$, in order to mimic the observed random behavior of the classification. Then the probability of misclassifying the object will be equal to $1 - p(c_i)$ and the expected misclassification probability is

$$P_e = \sum_{i=1}^{m} p(c_i)(1 - p(c_i)) = \frac{H_C^2}{2}. \quad (34)$$

Thus reducing the Gini index amounts to reducing the misclassification error associated with a randomized classification. The Gini index is also equal to the variance of the class-indicator regression variable (defined by $y_i(o) = 1$ if $c(o) = c_i$, and $y_i(o) = 0$ otherwise). Thus, reducing the Gini index consists also of reducing the residual variance of class indicator variables.

From the preceding discussion it follows also that the expected value of the quadratic entropy conditioned on the attribute values is equal to twice the asymptotic error rate of the nearest neighbor rule.

3.1 Quadratic conditional entropies and information

As in §2.1, the conditional entropy of c is defined by

$$H^2_{C|T} \triangleq \sum_{j=1}^{k} p(t_j) H^2_{C|t_j} = 1 - \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{p^2(c_i, t_j)}{p(t_j)}. \quad (35)$$

The quadratic information provided by t on c is defined by

$${I^2}^T_C \triangleq H^2_C - H^2_{C|T}. \quad (36)$$

Similarly, one may define

$$H^2_{C|T} \triangleq \sum_{i=1}^{m} p(c_i) H^2_{T|c_i} = 1 - \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{p^2(c_i, t_j)}{p(c_i)}. \quad (37)$$

The quadratic information provided by c on t is defined by

$${I^2}^C_T \triangleq H^2_T - H^2_{T|C}. \quad (38)$$

It is worth noting that in general ${I^2}^T_C \neq {I^2}^C_T$.

In the CART method, Breiman et al. use ${I^2}^T_C$ as an attribute selection criterion [12]. Given the very similar behavior of quadratic and logarithmic entropies, this criterion must admittedly suffer from similar difficulties than the logarithmic information criterion of §2.2. In particular, it favors many-valued splits and makes the comparison of scores for different values of the prior entropy difficult.

3.2 Normalizations

We are not surprised that the same normalization "medicine" has been applied to derive from the quadratic entropy an appropriate optimal splitting criterion. We will merely indicate the definition of the resulting *symmetrical* τ measure proposed in [17],

$$\tau \triangleq \frac{{I^2}^T_C + {I^2}^C_T}{H^2_T + H^2_C}, \quad (39)$$

which is the exact equivalent of our own C^T_C measure.

Of course the advantages of the latter measure are the same than those of C^T_C, no more no less.

3.3 Hypothesis testing

In the second part of their paper the authors of [17] present the use of an associated $\chi-$square hypothesis test. They note indeed that the quantities

$$(n_{..} - 1)(m-1)\frac{{I^2}^T_C}{H^2_C} \text{ or } (n_{..} - 1)(k-1)\frac{{I^2}^C_T}{H^2_T} \quad (40)$$

are distributed according to a $\chi-$square law with $(m-1)(k-1)$ degrees of freedom.

4 Other Loss and Distance Functions

Many other criteria have been proposed in various decision tree induction algorithms (see e.g. [13] for an interesting discussion of general divergence measures and their algorithmic properties).

For example to avoid bias towards many valued splits and overfitting problems, one approach consists of using modified estimates of relative frequencies such as

$$\hat{p}_i = \frac{n_i + \lambda}{n + m\lambda} \quad \forall\, i = 1, \ldots, m, \quad (41)$$

λ being a problem dependent parameter [18, 19, 20].

Let us briefly describe the **Kolmogorov-Smirnov** criterion proposed by Friedman [21] as an attribute selection criterion in decision tree induction, and afterwards extended for pruning [22]. The basic method is restricted to the two-class case and to ordered (e.g. numerically continuous) attributes.

Denoting by $F_{c_1}(a_i)$ (resp. $F_{c_2}(a_i)$) the (cumulative) probability distribution of an attribute conditioned to class c_1 (resp. c_2), the Kolmogorov-Smirnov distance is defined by

$$D^{KS}_C(a_i^*) = \max_{a_i} |F_{c_1}(a_i) - F_{c_2}(a_i)|. \quad (42)$$

The sampling distribution of D^{KS}_C under the independence assumption (i.e. if $F_{c_1} = F_{c_2}$) is independent of the distribution $F(a_i)$, yielding thus a non-parametric hypothesis test of the independence of a_i and c.

Note that the sampling distribution (and thus the levels of significance) depends on the sample sizes of each class which are however constant at a given tree node and independent of the considered attribute. Thus the ranking of $D^{KS}_C(a^*)$ is equivalent to the ranking of the significance levels, and the optimal splitting rule derived by Friedman consists of splitting a node by the attribute a_* corresponding to the maximum Kolmogorov-Smirnov distance,

$$D^{KS}_C(a_*^*) = \max_i D^{KS}_C(a_i^*), \quad (43)$$

together with its optimal threshold a^*_*.

The corresponding stop-splitting rule consists of checking that the significance level $1 - \alpha$ corresponding to $D(a^*_*)$ is smaller than a fixed threshold [22].

To appraise this criterion, we have applied it to our power system security problem. The corresponding variation of the sample values of $F_{\text{Sec}}(\text{Trbj})$, $F_{\text{Insec}}(\text{Trbj})$ and $D^{KS}_C(\text{Trbj})$ are illustrated in Fig. 5.

We note that the overall shape of the D^{KS}_C curve is quite similar to the shape of the curves in Fig. 4. It reaches its maximum value at 7310.5MW, which is very close to the maximum of 7308.5MW of curves C^T_C and D^T_C of Fig. 4. The behavior of D^{KS}_C is however smoother than the other measures, suggesting that its optimum threshold may be less sensitive to sampling noise.

5 Convexity or no convexity ?

Much of the above discussion turns around the convexity property of uncertainty measures, which has been often quoted as a desirable, if not necessary property in the context of inducing classification or class-probability trees.

However, if we want to reduce the tendency of favoring many valued splits - and incumbent overfitting problems - we should look for non-convex score measures.

Another situation where the convexity property leads to undesirable consequences is in the context of fuzzy tree induction where it yields a systematic bias in favor of crisp discriminators [23, 24, 25], instead of fuzzy ones. In this case, it prevents one from realizing the necessary compromise between model smoothness and information quantity. Again, it is necessary to deconvexify the measure by normalizing it by an appropriate regularization term [26].

We conclude that while convexity is a nice structural property which may be exploited to improve computational performances, it may also lead to undesirable results, in particular increased variance and hence overfitting. In such circumstances, a natural way to reduce the variance and overfitting is to deconvexify the measure.

References

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

[2] P. A. Devijver (1976). Entropie quadratique et reconnaissance de formes. In J. Simon, editor, *NATO ASI Series, Computer Oriented Learning Processes*. Noordhoff, Leyden.

[3] L. Wehenkel (1990). *Une approche de l'intelligence artificielle appliquée à l'évaluation de la stabilité transitoire des réseaux électriques*. PhD thesis, University of Liège. In French.

[4] M. D. Richard and R. P. Lippmann (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, **3**:461–483.

[5] L. Wehenkel (1993). Decision tree pruning using an additive information quality measure. In B. Bouchon-Meunier, L. Valverde and R. Yager, editors, *Uncertainty in Intelligent Systems*, 397–411. Elsevier - North Holland.

[6] S. Guiasu (1993). A unitary treatment of several known measures of uncertainty induced by probability, possibility, fuzziness, plausibility and belief. In B. Bouchon-Meunier, L. Valverde and R. Yager, editors, *Uncertainty in Intelligent Systems*, 355–365. Elsevier - North Holland.

[7] R. L. de Mántaras (1991). A distance-based attributes selection measure for decision tree induction. *Machine Learning*, **6**:81–92. Technical Note.

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

[9] T. O. Kvålseth (1987). Entropy and correlation: Some comments. *IEEE Trans. on Systems, Man and Cybernetics*, **SMC-17**(3):517–519.

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

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

[12] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth International (California).

[13] P. A. Chou (1991). Optimal partitioning for classification and regression trees. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **PAMI-13**(14):340–354.

[14] L. Wehenkel, T. Van Cutsem and M. Ribbens-Pavella (1989). An artificial intelligence framework for on-line transient stability assessment of power systems. *IEEE Trans. on Power Syst.*, **PWRS-4**:789–800.

[15] D. Michie, D. Spiegelhalter and C. Taylor, editors, (1994). *Machine learning, neural and statistical classification*. Ellis Horwood. Final rep. of ESPRIT project 5170 - StatLog.

[16] L. Wehenkel, I. Houben, M. Pavella, L. Riverin and G. Versailles (1995). Automatic learning approaches for on-line transient stability preventive control of the Hydro-Québec system - Part I. Decision tree approaches. In *Proc. of SIPOWER'95, 2nd IFAC Symp. on Control of Power Plants and Power Systems*, 231–236.

[17] X. Zhou and T. S. Dillon (1991). A statistical-heuristic feature selection criterion for decision tree induction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **PAMI-13**:834–841.

[18] W. L. Buntine (1990). *A theory of learning classification rules*. PhD thesis, School of Computing Science, Sidney University of Technology.

[19] J. R. Quinlan (1987). Simplifying decision trees. *Int. J. of Man-Mach. Studies*, **27**:221–234.

[20] A. Zighed, J. Auray and G. Duru (1992). *SIPINA. Méthode et Logiciel*. Alexandre Lacassagne - Lyon.

[21] J. H. Friedman (1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Trans. on Computers*, **C-26**:404–408.

[22] E. M. Rounds (1980). A combined nonparametric approach to feature selection and binary decision tree design. *Pattern Recognition*, **12**:313–317.

[23] X. Boyen and L. Wehenkel (1995). On the unfairness of convex discrimination quality measures for fuzzy partitioning in machine learning. Technical report, University of Liège.

[24] X. Boyen and L. Wehenkel (1995). Automatic induction of continuous decision trees. To appear in *Proc. of IPMU96, Info. Proc. and Manag. of Uncertainty in Knowledge-Based Systems*, Granada (SP).

[25] M. Ramdani (1994). *Système d'Induction Formelle à base de Connaissances Imprécises*. PhD thesis, Univ. Paris VI.

[26] X. Boyen and L. Wehenkel (1995). Fuzzy decision tree induction for power system security assessment. In *Proc. of SIPOWER'95, 2nd IFAC Symp. on Control of Power Plants and Power Systems*, 151–156, Mexico.