Entropy evaluation based on confidence intervals of frequency estimates : Application to the learning of decision trees

Mathieu Serrurier SERRURIER@IRIT.FR
IRIT - Université Paul Sabatier 118 route de Narbonne 31062, Toulouse Cedex 9, France

Henri Prade PRADE@IRIT.FR
IRIT - Université Paul Sabatier, Toulouse, France & QCIS, University of Technology, Sydney, Australia

Abstract

Entropy gain is widely used for learning decision trees. However, as we go deeper downward the tree, the examples become rarer and the faithfulness of entropy decreases. Thus, misleading choices and over-fitting may occur and the tree has to be adjusted by using an early-stop criterion or post pruning algorithms. However, these methods still depends on the choices previously made, which may be unsatisfactory. We propose a new cumulative entropy function based on confidence intervals on frequency estimates that together considers the entropy of the probability distribution and the uncertainty around the estimation of its parameters. This function takes advantage of the ability of a possibility distribution to upper bound a family of probabilities previously estimated from a limited set of examples and of the link between possibilistic specificity order and entropy. The proposed measure has several advantages over the classical one. It performs significant choices of split and provides a statistically relevant stopping criterion that allows the learning of trees whose size is well-suited w.r.t. the available data. On the top of that, it also provides a reasonable estimator of the performances of a decision tree. Finally, we show that it can be used for designing a simple and efficient online learning algorithm.

1. Introduction

Although decision tree methods have been one of the first machine learning approaches, they remain popular because of their simplicity and flexibility. Most algorithms for building decision trees are based on the use of information gain function for choosing the best attribute for splitting the data at each step of the learning process. Thus, the ID3 algorithm is based on logarithmic entropy (Quinlan, 1986), while CART (Breiman et al., 1984) is based on the Gini impurity measure. Numerous alternatives have been proposed for the gain function (Buntine & Niblett, 1992; Lee, 2001; Nowozin, 2012). However, one drawback in this kind of approach is that the gain function becomes less and less significant when the number of examples in the considered node decreases. In the particular case of log entropy-based gain, which is still one of the most largely used, splitting a node always decreases the weighted entropy of the leaves obtained. It then leads to learn trees that may overfit the data and then decreases the performance of the algorithm. This can be avoided by using early-stop criterion or post-pruning methods. However, these methods still depend on the initial choices based on the entropy calculus, even if this evaluation may be not significant. The main limitation of the log-based entropy (but this also applies to some extent to its multiple refinements) is that it does not take into account the amount of data used for estimating the frequencies on the different classes.

The goal of this paper is to show how to extend the classical entropy calculus in order to take into account the amount of information available and then having a single entropy measure that addresses the different issues of the decision tree learning process in an elegant way. We propose to use the upper bound of the frequency estimates for defining a so-called possibilistic cumulative entropy. The approach relies on the building of a possibility distribution. Quantitative possibility measures can be viewed as upper bounds of probabilities. Then, a possibility distribution represents a family of probability distributions (Dubois & Prade, 1992). In agreement with this view, a probability-possibility transformation has

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

been proposed (Dubois et al., 1993). This transformation associates a probability distribution with the maximally specific (restrictive) possibility distribution which is such that the possibility of any event is indeed an upper bound of the corresponding probability. Possibility distributions are then able to describe epistemic uncertainty and to represent knowledge states such as total ignorance, partial ignorance or complete knowledge. Starting from the link between the specificity order over possibility distribution and the entropy of a probability distribution, we propose a log-based loss function for possibility distributions based on (Serrurier & Prade, 2013). We derive the possibilistic cumulative entropy function for a possibility distribution associated to a frequency distribution. Then, we build a possibility distribution that upper bounds the confidence intervals of the frequency values (according to the number of data available and a confidence degree) and we compute its relative possibilistic entropy. This cumulative entropy has nice properties. For instance, it respects the entropy order for a fixed level of information and this entropy increases for a fixed frequency distribution when the amount of data decreases. It also provides a stopping criterion when splitting nodes is no longer significant. Thus, it allows to choose the most relevant nodes instead of reasoning a posteriori about the significance of the choices made on the basis of classic entropy, as done with early stop criteria or post-prunning methods (see (Esposito et al., 1997) for a review of pruning methods). Thanks to this ability, we propose a direct extension of the classical algorithm with possibilistic entropy and we show how to easily extend it to obtain an incremental online algorithm. Last, possibilistic cumulative entropy also provides a global evaluation measure of a decision tree that is a relevant estimation of its performances outside the training set.

The paper is organized as follows. First we provide a short background on possibility distributions and possibility measures and their use as upper bounds of families of probability distributions. Second, we describe possibilistic cumulative entropy with its properties. Section 4 is devoted to the presentation of the two algorithms and their comparisons with state of the art approaches. As our goal is to demonstrate the benefits of our measure with respect to classical log entropy, we compare the performances of these approaches on 16 benchmark databases in the last section.

2. Possibility theory

Possibility theory, introduced in (Zadeh, 1978), was initially proposed in order to deal with imprecision and uncertainty due to incomplete information, as the one provided by linguistic statements. This kind of epistemic uncertainty cannot be handled by a single probability distribution, especially when a priori knowledge about the nature of the probability distribution is lacking. A possibility distribution π on a discrete universe $\Omega = \{c_1, \dots, c_q\}$ is a mapping from Ω to $[0, 1]$. We note $\mathit{\Pi}$ the set of all possibility distributions over Ω. The value $\pi(c)$ is called the possibility degree of the value c in Ω. For any subset of Ω, the possibility measure is defined as follows:

$$\forall A \subseteq \Omega, \Pi(A) = \sup_{c \in A} \pi(c).$$

If it exists at least a value $c \in \Omega$ for which we have $\pi(c) = 1$, the distribution is normalized. One view of possibility theory is to consider a possibility distribution as a family of probability distributions (see (Dubois, 2006) for an overview). Thus, a possibility distribution π will represent the family of the probability distributions for which the measure of each subset of Ω will be respectively lower and upper bounded by its necessity and its possibility measures. More formally, if $\mathcal{P}$ is the set of all probability distributions defined on Ω, the family of probability distributions $\mathcal{P}(\pi)$ associated with π is defined as follows:

$$\mathcal{P}(\pi) = \{p \in \mathcal{P}, \forall A \in \Omega, P(A) \leq \Pi(A)\}. \quad (1)$$

where P is the probability measure associated with p. We can distinguish two extreme cases of information situations: i) *complete knowledge* $\exists c \in \Omega$ such as $\pi(c) = 1$ and $\forall c' \in \Omega, c' \neq c, \pi(c) = 0$ and ii) *total ignorance* (i.e. $\forall c \in \Omega, \pi(c) = 1$) that corresponds to the case where all probability distributions are possible. This type of ignorance cannot be described by a single probability distribution. According to this probabilistic interpretation, a method for transforming probability distributions into possibility distributions has been proposed in (Dubois et al., 1993). The idea behind this is to choose the most informative possibility measure that upper bounds the considered probability measure. We note S_q the set of permutations of the set $\{1, \dots, q\}$. We introduce the notion of σ-specificity which is a partial pre-order:

Definition 1 (σ-specificity) *The distribution π is more σ-specific than π', denoted $\pi \preceq_\sigma \pi'$, if and only if :*

$$\pi \preceq_\sigma \pi' \Leftrightarrow \exists \sigma \in S_q, \forall i \in \{1, \dots, q\}, \pi(c_i) \leq \pi'(c_{\sigma(i)}) \quad (2)$$

Then, the possibility measure obtained by probability-possibility transformation corresponds to the most σ specific possibility distribution which bounds the distribution. We denote T_p^* the possibility distribution obtained from p by the probability-possibility transformation. This distribution has the following property:

$$\forall \pi, p \in \mathcal{P}(\pi) \Rightarrow T_p^* \preceq_\sigma \pi. \quad (3)$$

For each permutation $\sigma \in S_q$ we can build a possibility distribution T_p^σ which encodes p as follows:

$$\forall j \in \{1,\ldots,q\}, T_p^\sigma(c_j) = \sum_{k,\sigma(k)\leq\sigma(j)} p(c_k). \qquad (4)$$

Then, each T_p^σ corresponds to a cumulative distribution of p according to the order defined by σ. We have:

$$\forall \sigma \in S_q, p \in \mathcal{P}(T_p^\sigma)$$

The probability-possibility transformation (Dubois et al., 2004) (noted P-Π transformation) uses one of these particular possibility distributions.

Definition 2 (P-Π transformation (discrete case)) *Given a probability distribution p on $\Omega = \{c_1,\ldots,c_q\}$ and a permutation $\sigma^* \in S_q$ such as $p(c_{\sigma^*(1)}) \leq \ldots \leq p(c_{\sigma^*(q)})$, the P-Π transformation of p is noted T_p^* and is defined as:*

$$T_p^* = T_p^{\sigma^*}.$$

T_p^* is the cumulative distribution of p built by considering the increasing order of p. For this order, T_p^* is the most specific possibility distribution that encodes p. We have then the following properties

$$\forall \sigma \in S_q, T_p^* \preceq_\sigma T_p^\sigma. \qquad (5)$$

Example 1 *For instance, we consider p on $\Omega = \{c_1,c_2,c_3\}$ with $p(c_1) = 0.5$, $p(c_2) = 0.2$ and $p(c_3) = 0.3$. We obtain $\sigma^*(1) = 3$, $\sigma^*(2) = 1$, $\sigma^*(3) = 2$ and then $T_p^*(c_1) = 0.5 + 0.3 + 0.2 = 1$, $T_p^*(c_2) = 0.2$ and $T_p^*(c_3) = 0.3 + 0.2 = 0.5$.*

The interest of comparing the entropy of probability distribution by considering the σ-specificity order of its P-Π transformation has been emphasized in (Dubois & Hüllermeier, 2007) with the following key property :

$$\forall p, p' \in \mathcal{P}, T_p^* \preceq_\sigma T_{p'}^* \Rightarrow \mathcal{H}(p) \leq \mathcal{H}(p') \qquad (6)$$

where $\mathcal{H}(p)$ is an entropy function.

3. Possibilistic cumulative entropy

We now explain how particular possibility distributions can be used to take into account the amount of data used for estimating the frequencies in the computation of the entropy.

3.1. Probabilistic loss function and entropy

Probabilistic loss functions are used for evaluating the differences between a probability distribution with respect to data. In particular, we look for concave loss function $\mathcal{L}(f,X)$ which is linear w.r.t. $X = \{x_1,\ldots,x_n\}$, i.e. $\mathcal{L}(f,X) = \frac{\sum_{i=1}^n \mathcal{L}(f,x_i)}{n}$, and where f is a distribution (probabilistic or possibilistic). Let $\alpha_1,\ldots,\alpha_q$ be the frequency of the elements of X that belong respectively to $\{c_1,\ldots,c_q\}$. We note

$$\mathbb{1}_j(x_i) = \begin{cases} 1 & \text{if } x_i = c_j \\ 0 & \text{otherwise.} \end{cases}$$

The logarithmic-based likelihood is defined as follows:

$$\mathcal{L}_{log}(p|x_i) = -\sum_{j=1}^{q} \mathbb{1}_j(x_i) log(p_j). \qquad (7)$$

When we consider the whole set of data we obtain:

$$\mathcal{L}_{log}(p|X) = -\sum_{j=1}^{q} \alpha_j log(p_j). \qquad (8)$$

When p is estimated with respect to frequencies, we obtain the entropy of the distribution (which corresponds to the minimum of the loss function).

$$\mathcal{H}(p) = -\sum_{j=1}^{q} p_j log(p_j). \qquad (9)$$

The higher the entropy, the lower the amount of information (uniform distribution). The entropy is equal to 0 when the probability is equal to 1 for one class. Entropy is the basis of the majority of algorithms for learning decision trees. The goal is to build a decision tree for which each leaf describes a probability over class with the lowest possible entropy.

3.2. Possibilistic loss function and entropy

In this section we show how to use $\mathcal{L}_{log}$ in order to define a loss function, and the related entropy, for possibility distributions that agrees with the interpretation of a possibility distribution in terms of a family of probability distributions. Proofs and detailed discussion about possibilistic loss function can be found in (Serrurier & Prade, 2013). We expect four properties:

(a) The possibilistic loss function is minimal for the possibility distribution that results from the P-Π transformation of the frequencies.

(b) As for probability distribution, the possibilistic entropy will be a linear function of possibilistic loss function applied to a set of data X_p that supports a probability distribution p.

(c) The possibilistic entropy applied to P-Π transformations respects the specificity order as in (6).

(d) The possibilistic entropy increases when uncertainty around the considered probability distribution increases.

Since a possibility distribution π can be viewed as an upper bound of a cumulative function, for all j, the pair $\pi_j = (\pi(c_{\sigma(j)}), 1 - \pi(c_{\sigma(j)}))$ (σ is the permutation of S_q such that $\pi(c_{\sigma(1)}) \leq \ldots \leq \pi(c_{\sigma(q)})$) can be seen as a Bernouilli probability distribution for the sets of events $BC_j = \bigcup_{i=1}^{j} c_{\sigma(i)}$ and $\overline{BC_j}$. Then, the logarithmic loss of a possibility distribution for an event will be the average of the log loss of each binomial distribution π_j re-scaled in $[0, 0.5]$ where the entropy function $-x * log(x) - (1 - x) * log(1 - x)$ is strictly increasing. This re-scaling is necessary for having proposition 1 and 2 below.

$$\begin{aligned}&\mathcal{L}_{\pi\text{-}l}(\pi|X)) = \\ &-\sum_{j=1}^{q}(\frac{cdf_j}{2} * log(\frac{\pi_j}{2}) + (1 - \frac{cdf_j}{2}) * log(1 - \frac{\pi_j}{2})).\end{aligned} \tag{10}$$

where $cdf_j = \sum_{k,\sigma(k)\leq\sigma(j)} \alpha_k$. If we only consider one piece of data x such that $x \in c_j$ we obtain :

$$\begin{aligned}&\mathcal{L}_{\pi\text{-}l}(\pi|x) = -\sum_{j=1}^{q}(log(1 - \frac{\pi_j}{2})) \\ &-\frac{1}{2} * \sum_{i,\sigma(i)\geq\sigma(j)}^{q}(log(\frac{\pi_{\sigma(i)}}{2}) - log(1 - \frac{\pi_{\sigma(i)}}{2})).\end{aligned} \tag{11}$$

It can be checked that this loss function is indeed linear w.r.t. X. The property (a) has been proven in (Serrurier & Prade, 2013). We remark that cdf_j corresponds to the cumulative probability distribution of the frequencies with respect to σ (Eq. 4). Then, we can derive a definition of the entropy of a possibility distribution relative to a probability distribution:

$$\begin{aligned}\mathcal{H}_{\pi\text{-}l}(p, \pi) &= \frac{\mathcal{L}_{\pi\text{-}l}(\pi|X_p)}{q * log(q)} \\ &= -\sum_{j=1}^{q}\frac{\frac{T^*_p(c_j)}{2} * log(\frac{\pi(c_j)}{2}) + (1 - \frac{T^*_p(c_j)}{2}) * log(1 - \frac{\pi(c_j)}{2})}{q * log(q)}.\end{aligned} \tag{12}$$

where X_p is a set of data that supports a probability distribution p. $q * log(q)$ is a normalization factor. The expected property (b) is obvious if we consider the probability distribution p such as $p(c_i) = \alpha_i$. We can now establish some properties of the possibilistic entropy.

Proposition 1 *Given two probability distributions p and p' on $\Omega = \{c_1, \ldots, c_q\}$ we have:*

$$T^*_p \preceq T^*_{p'} \Rightarrow \mathcal{H}_{\pi\text{-}l}(p, T^*_p) \leq \mathcal{H}_{\pi\text{-}l}(p', T^*_{p'})$$

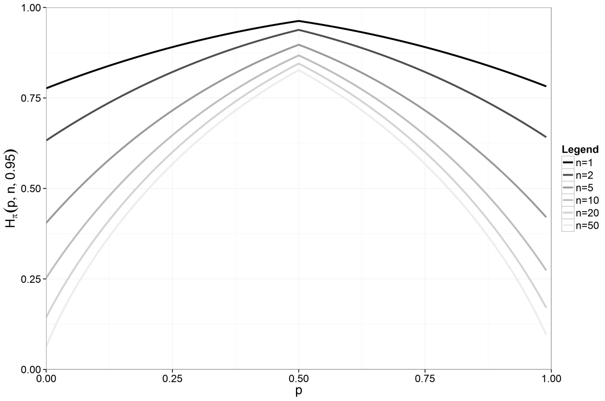

Figure 1. Possibilistic cumulative function of a binomial probability distribution on $\Omega = \{c_1, c_2\}$ with $\gamma = 0.05$ for different values of n. The x-axis represents the value of $p(c_1)$ and the y-axis the value $\mathcal{H}^*_{\pi\text{-}l}(p, n, 0.05)$.

Proof (sketch) We can assume without loss of generality that the values of distributions p and p' are in increasing order. It can be easily shown that the re-scaled entropy of the binomial counterpart of p restricted to the events BC_j and $\overline{BC_j}$ is less than the entropy of the binomial counterpart of p' on the same events.

Proposition 2 *Given a probability distribution p and two possibility distributions π and π' on $\Omega = \{c_1, \ldots, c_q\}$ we have:*

$$T^*_p \preceq \pi \preceq \pi' \Rightarrow \mathcal{H}_{\pi\text{-}l}(p, T^*_p) \leq \mathcal{H}_{\pi\text{-}l}(p, \pi) \leq \mathcal{H}_{\pi\text{-}l}(p, \pi')$$

Proof This property is naturally obtained from the definitions of $\mathcal{H}_{\pi\text{-}l}$ and the previous.

These two last propositions validate the properties (c) and (d) and show that the possibility cumulative entropy can be used for measuring both the entropy and the epistemic uncertainty and is fully compatible with the interpretation of a possibility distribution as a family of probability distributions. We can also notice that possibilistic cumulative entropy is equal to 0 for *complete knowledge* (as for classical entropy) and equal to 1 for *total ignorance* (and not for uniform distributions, as for classical entropy).

3.3. Possibilistic cumulative entropy of a frequency distribution

As said previously, the entropy calculus does not take into account the amount of information used for estimating the frequencies. The idea behind possibilistic cumulative entropy is to consider the confidence intervals around the estimation of the frequencies to have an entropy measure that increases when the size of the confidence interval increases.

Applying directly the entropy to the upper-bounds of the frequency is not satisfactory since entropy only applies to genuine probability distribution. We propose to build the most specific possibility distribution that upper bounds the confidence interval and compute its possibilistic entropy relative to the frequency distribution.

We use the Agresti-Coull interval (see (Agresti & Coull, 1998) for a review of confidence intervals for binomial distributions) for computing the upper bound value of the probability of an event. Given $p(c)$ the probability of the event c estimated from n pieces of data, the upper bound $p^*_{\gamma,n}$ of the $(1-\gamma)\%$ confidence interval of p is obtained as follows:

$$p^*_{\gamma,n}(c) = \tilde{p} + z\sqrt{\frac{1}{\tilde{n}}\tilde{p}(1-\tilde{p})} \qquad (13)$$

where $\tilde{n} = n + z^2$, $\tilde{p} = \frac{1}{\tilde{n}}(p(c) * n + \frac{1}{2}z^2)$, and z is the $1 - \frac{1}{2}\gamma$ percentile of a standard normal distribution. The most specific possibility distribution $\pi^\gamma_{p,n}$ that contains upper bounds of the $(1-\gamma)\%$ confidence interval of p estimated from n piece of data is computed as follows:

$$\pi^\gamma_{p,n}(c_j) = P^*_{\gamma,n}(\bigcup_{i=1}^{j}\{c_{\sigma(i)}\}) \qquad (14)$$

where $\sigma \in S_q$ is the permutation such as $p(c_{\sigma(1)}) \leq \ldots \leq p(c_{\sigma(q)})$. Thus, $\pi^\gamma_{p,n}$ is built in the same way as π^*_p except that it also takes into account the uncertainty around the estimation of p. Obviously, we have $p \in \mathcal{P}(\pi^\gamma_{p,n})$, $\forall n > 0, \pi^*_p \preceq \pi^\gamma_{p,n}$ and $\lim_{n\to\infty} \pi^\gamma_{p,n} = \pi^*_p$. Having $\pi^\gamma_{p,n}$, we can now define the possibilistic cumulative entropy of a probability distribution:

$$\mathcal{H}^*_{\pi\text{-}l}(p, n, \gamma) = \mathcal{H}_{\pi\text{-}l}(p, \pi^\gamma_{p,n}) \qquad (15)$$

Fig. 1 illustrates the different values of $\mathcal{H}^*_{\pi\text{-}l}$ for a binomial distribution with different values of n. We can observe that the value of $\mathcal{H}^*_{\pi\text{-}l}$ increases when n decreases for the same distribution.

Proposition 3 *Given a probability distribution* p *on* $\Omega = \{c_1, \ldots, c_q\}$ *and* $n' \leq n$ *we have:*

$$\forall\gamma \in]0, 1[, \mathcal{H}^*_{\pi\text{-}l}(p, n, \gamma) \leq \mathcal{H}^*_{\pi\text{-}l}(p, n', \gamma)$$

Proof Use the property $\pi^*_p \preceq \pi^\gamma_{p,n} \preceq \pi^\gamma_{p,n'}$ and proposition 2.

Proposition 4 *Given two probability distributions* p *and* p' *on* $\Omega = \{c_1, \ldots, c_q\}$ *we have:*

$$\forall\gamma \in]0, 1[, T^*_p \preceq T^*_{p'} \Rightarrow \mathcal{H}^*_{\pi\text{-}l}(p, n, \gamma) \leq \mathcal{H}^*_{\pi\text{-}l}(p', n, \gamma)$$

Proof (sketch) use the same as proposition 1 and use the property $p(c) \leq p'(c) \Rightarrow p^*_{\gamma,n}(c) \leq p^{'*}_{\gamma,n}(c)$.

These two last propositions show that possibilistic cumulative entropy has the expected properties and can take effectively into account the uncertainty around the estimation of the frequency distribution.

Example 2 *We consider* p *on* $\Omega = \{c_1, c_2, c_3\}$ *with* $p(c_1) = 0.5$, $p(c_2) = 0.2$ *and* $p(c_3) = 0.3$. *For* $n = 10$ *and* $\gamma = 0.05$. $\pi^{0.05}_{p,10}(c_1) = P^*_{0.05,10}(\{c_1, c_2, c_3\}) = 1$, $\pi^{0.05}_{p,10}(c_2) = p^*_{0.05,10}(c_2) = 0.52$, $\pi^{0.05}_{p,10}(c_3) = P^*_{0.05,10}(\{c_2, c_3\}) = 0.76$ *and* $\mathcal{H}^*_{\pi\text{-}l}(p, 10, 0.05) = 0.81$.

4. Learning decision trees with possibilistic cumulative entropy

In this section, we propose two different algorithms that are based on the possibilistic cumulative entropy. The first one is the classical decision tree learning algorithm for which the gain function is now based on possibilistic cumulative entropy. In the next subsection we show that the possibilistic cumulative entropy can be used for revising a decision tree and then we obtain an incremental decision tree algorithm.

4.1. Possibilistic cumulative entropy for decision trees

The building of a decision tree is based on the recursive induction of the nodes. For learning a node, the best attribute according to the gain is chosen. Given a set Z of n examples and an attribute A (real valued attributes are handled by means of binary attributes associated with thresholds) which has $v1, \ldots, vr$ possible values. We note p_Z the probability distribution of the classes for the examples in Z, p_{vk} the probability distribution of the classes for the examples for which the value of A is vk and $|vk|$ the size of this set. The classical gain function is

$$G(Z, A) = \mathcal{H}(p_Z) - \sum_{k=1}^{r} \frac{|vk|}{n}\mathcal{H}(p_{vk}). \qquad (16)$$

As pointed into the introduction, this approach suffers some major drawbacks. First, the gain is always positive and can not be used as a stop criterion. The idea behind this is that splitting the set of examples always decreases the entropy. However, when going deeper and deeper in the tree, less and less examples are used for computing the entropy and the result may not be significant. Moreover, the gain tends to favor the nominal attributes having a lot of possible values. In this paper we propose to use a new gain based on possibilistic cumulative entropy. Since it takes into account the uncertainty around the estimation of the probability distribution, the gain can be negative if the splitting has no statistically significant advantage. The gain function we propose is defined as follows:

$$G^{\pi}_{\gamma}(Z,A) = \mathcal{H}^{*}_{\pi\text{-}l}(p_Z, n, \gamma) - \sum_{k=1}^{r} \frac{|vk|}{n} \mathcal{H}^{*}_{\pi\text{-}l}(p_{vk}, |vk|, DS(\gamma, r)). \tag{17}$$

where $DS(\gamma, r)$ is the Dunn−Šidàk correction of γ for r comparison. By using G^{π}_{γ}, we have a faithful stop criterion and we penalize the attributes having a lot of possible values. G^{π}_{γ} also favors well-balanced trees where the number of examples in the nodes is significant enough for entropy computation. This approach directly produces trees that limit overfitting. The possibilistic cumulative entropy can also be used as a quality measure of a tree T with a set of leaves L :

$$\mathcal{H}^{*}_{\pi\text{-}l}(T, \gamma) = \sum_{l \in L} \mathcal{H}^{*}_{\pi\text{-}l}(p_l, n_l, \gamma) \tag{18}$$

where p_l is the frequency distribution of n_l training examples that fall in leaf l. The only parameter of the algorithm is γ. It represents the strength of the constraint for splitting node. This parameter has been tuned by choosing the best value of γ inside a set of 10 possible values by the mean of a cross-validation on the training set.

4.2. Online decision trees

A remarkable property of the possibilistic cumulative entropy and the associated gain function it that they can easily be used for revising a decision tree. We assume that the tree saves the set of the related examples for each leaf. The revision process for a new example x is the following:

1. browse recursively the tree to the corresponding leaf
2. add x to the set of examples
3. search the attribute with the best G^{π}_{γ}
4. if the gain is positive, create a new node with the corresponding attribute, else do nothing.

Since G^{π}_{γ} is positive if and only if we have enough data for performing a split of the node which can increase the learning power of the tree, the tree will grow up slowly. We can reasonably suppose that it exists an upper bound of the number of example N_{max} before which a node is always split since the size of the confidence interval decreases quickly when the number of example increases. In this case, the complexity of the revision of the tree with one example will be $\mathcal{O}(NB_A * N_{max} log(N_{max}))$ where NB_A is the number of attributes. The γ parameter is tuned as in the previous algorithm. Although it is not completely satisfactory in a genuine online procedure, it is not costly if it is done at the beginning of the algorithm. We can also imagine that the online procedure takes place on a repeated context.

4.3. State of the art

Some other approaches (see e.g. (Bernard, 2005)) have been proposed in order to consider the amount of data used for the evaluation the parameters of a probability distribution into the entropy calculus. The first one is to consider an apriori probability distribution (usually the uniform one) and to revise it with the observation. However, we can observe that the approach depends on the choice of the initial distribution and, since it is still an entropy computed on a single probability distribution, it does not make the difference between a uniform distribution obtained with a large number of estimations, and the initial distribution (if it is uniform).

Possibility distributions have already been used in machine learning for dealing with imprecise data (Jenhani et al., 2008), or for generalizing Ockham's razor principle when learning lazy decision trees (Hüllermeier, 2002). Our approach shares some ideas with the upper entropy proposed in (Abellàn & Moral, 2005). This works is based on the building of a belief function that encodes the confidence intervals around the estimation of a probability distribution. Then, the entropy computed is the maximal entropy of the probability distributions that are bounded by the belief function (with the optional addition of a term which corresponds to a non-specificity measure). However, there are some important differences with our work based on possibilistic cumulative entropy. First, due to the use of the maximum, the upper entropy is not a linear function of individual loss function (and then not a genuine entropy function). The second problem is that finding the maximum of entropy requires to use linear optimization algorithm which may be costly when the number of classes increases. The last difference comes from the use of the maximum. Indeed, when the number of the examples is small, the uniform probability distribution may be often in the confidence interval which prevents to make an informed choice since the entropy is equal to 1. In (Abellàn & Moral, 2005), the authors are led to restrict themselves to small confidence intervals (rather than faithful ones, as in our case) in order to avoid the previously mentioned pitfall.

ID3 and J4.8 use pessimistic error rate (based on confidence interval) as a pre or post pruning criteria. However, this is only a simple stopping or pruning criterion and it cannot be used for choosing attributes when splitting nodes. In (Nowozin, 2012), the author takes into consideration the numbers of examples in the parent node by using a refined version of the entropy (Grassberger, 1988). However, the gain is still always positive and the approach is less general than the one proposed in the current paper. Note that confidence intervals are used in (Katz et al., 2012) in the prediction of the class by taking into account the uncertainty on the values of the attributes, or on the split thresholds. On-

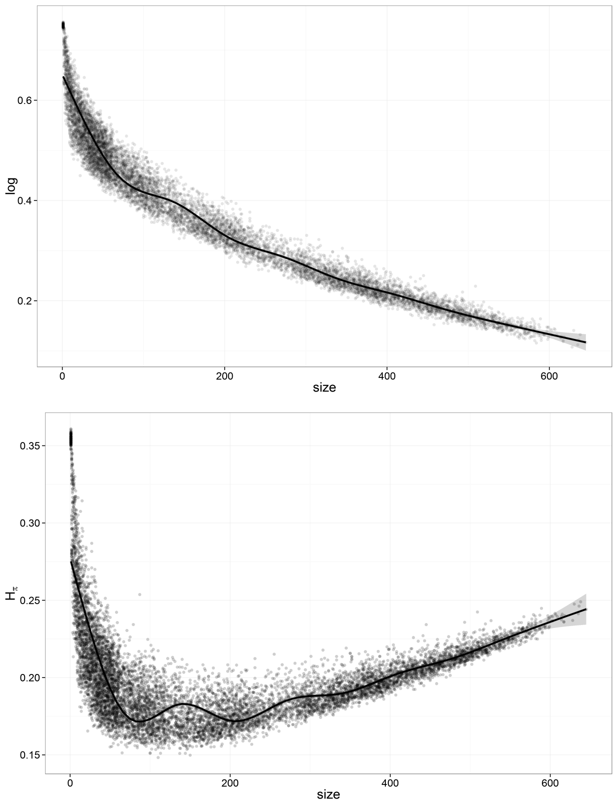

Figure 2. Entropy of the tree with respect to the size of the tree on the Yeast database. Classical entropy is on the top and $\mathcal{H}^*_{\pi\text{-}l}(T,\gamma)$ is on the bottom. Curves computed with LOESS

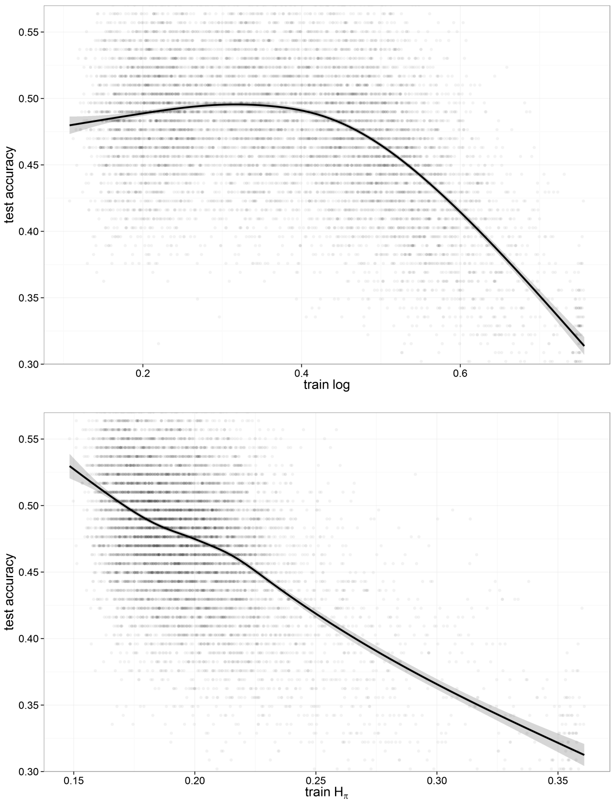

Figure 3. Accuracy of the tree on test set with respect to the entropy of the tree on the training set for the Yeast database. Classical entropy is on the top and $\mathcal{H}^*_{\pi\text{-}l}(T,\gamma)$ is on the bottom. Curves computed with LOESS

line algorithms have already been proposed in (Schlimmer & Fisher, 1986; Utgoff, 1989; Domingos & Geoff, 2000; Gama et al., 2006), but they are based on the revision of the *whole* tree with the new example and all (or a subset of) the previous examples.

5. Experiments

As pointed out in the introduction, the goal of the paper is to illustrate the interest of handling epistemic uncertainty in log-entropy calculus and to show the improvement w.r.t. the classical approach. We used 16 benchmarks from UCI[1]. 3 of these datasets have nominal attributes and 13 have numerical attributes only. We note ΠTree the decision tree learning algorithm based on the possibilistic cumulative entropy, O-ΠTree is its online counterpart. We compare them with the LogTree algorithm which is based on the same implementation as ΠTree, but which uses the log entropy (without post pruning). PrunTree is logTree with post pruning based on pessimistic error rate (the parameter γ for the confidence intervals has been tuned with the same method used for ΠTree) and J4.8 is the baseline (we use the Weka implementation with parameters tuned in a subset of data as for our approach) which uses more advanced pruning such as tree raising. Figures 2 and 3 illustrate the ability of $\mathcal{H}^*_{\pi\text{-}l}(T,\gamma)$ to provide meaningful trees on the Yeast database. The figures are obtained as follows: we split the database in a 90% (training)/10% (test) scheme, we generate 10 random trees of random sizes (i.e. attribute for a node is chosen randomly and the threshold is chosen alternatively with classical entropy, and with possibilistic cumulative entropy on the training set), we evaluate the entropy of the tree on the training set and its accuracy on the test set, we repeat this process 1000 times. Fig. 2 shows that the classical entropy of the tree always decreases when its size increases. In the case of $\mathcal{H}^*_{\pi\text{-}l}(T,\gamma)$, it shows that $\mathcal{H}^*_{\pi\text{-}l}(T,\gamma)$ first decreases with size and then increases when the tree becomes too complex w.r.t. the number of examples. Fig. 3 illustrates that it exists a threshold below which decreasing log entropy doesn't increase the accuracy (over fitting). On the contrary, decreasing $\mathcal{H}^*_{\pi\text{-}l}(T,\gamma)$ on the training set tends to increase the accuracy of the tree on the test set.

[1] http://www.ics.uci.edu/ mlearn/MLRepository.html

Data set	Log Tree	Pruned	PTree	O-PTree	J48
SOYBEAN	89.4±5.0	89.4±5.0	**94.0±2.8**	89.0±3.8	91.7±3.1
LYMPH	72.9±11.8	72.9±11.8	78.3±7.9	**78.3±8.2**	75.8±11.0
ZOO2	**97.0±4.8**	97.0±4.8	97.0±4.8	96.0±5.1	92.6±7.3
ILPD	67.9±5.5	67.4±5.6	**69.9±5.3**	66.8±4.7	69.3±6.3
YEAST	52.0±4.1	57.0±3.3	57.1±3.4	56.7±3.6	**57.8±5.5**
WAVEFORM	75.2±1.5	75.3±1.5	**77.4±1.5**	72.6±1.8	75.9±1.4
DIABETES	68.7±5.7	70.4±4.7	**74.3±4.4**	70.4±3.4	74.2±5.1
BANKNOTE	98.3±1.1	98.3±1.1	98.3±1.0	97.4±2.1	**98.5±1.0**
ECOLI	78.9±7.7	80.4±7.4	82.4±7.9	**83.6±7.2**	83.3±8.5
VEHICLE	71.6±4.7	71.6±4.0	**74.1±4.1**	69.1±3.1	73.3±5.0
IONOSPHERE	90.3±4.7	90.3±4.7	91.1±3.6	87.7±4.0	**91.4±3.7**
SEGMENT	96.8±0.6	96.7±0.7	96.9±1.2	94.7±1.4	**97.1±1.1**
PENDIGITS	**96.5±0.5**	96.4±0.5	96.4±0.2	93.2±1.0	96.5±0.5
SPAMBASE	91.8±1.2	91.7±1.3	**94.0±1.3**	90.5±1.2	92.9±1.0
BREAST-WV2	92.9±2.4	92.9±2.4	93.9±3.1	**94.7±1.6**	94.1±3.5
WINE2	92.5±8.7	92.5±8.7	93.7±7.3	**94.3±8.3**	94.1±3.5

Table 1. Classification accuracy of LogTree, PrunTree, ΠTree and O-ΠTree, J4.8 on different databases

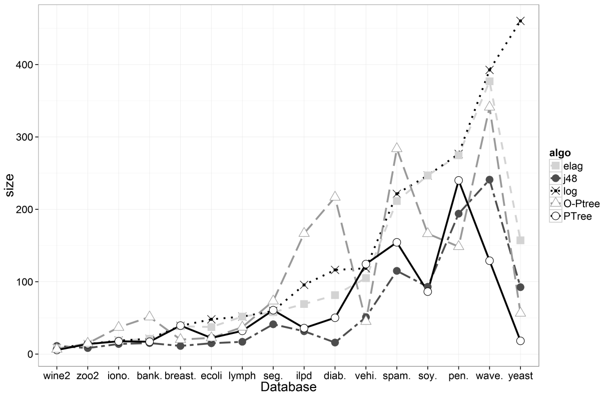

Figure 4. Number leaves of LogTree, PrunTree, ΠTree and O-ΠTree, J4.8 comparison for different databases.

Table 1 reports the accuracy results for the different databases. Highest values are in bold, underlined results indicate that the algorithm is statistically better (paired T-Test) than its two opponents (logTree, PrunTree vs O-ΠTree, ΠTree, J4.8 is not taken into account). We use a Wilcoxon signed ranked test (Demšar, 2006) for comparing the algorithms. ΠTree significantly outperforms its classical competitors (there is no significant statistical difference with J4.8). We do not observe significant difference between O-ΠTree and LogTree and PrunTree. This can be considered as a good result for an online algorithm.

Fig. 4 compares the number of leaves for the trees induced by the algorithms. As expected LogTree always produces the most complex trees. ΠTree algorithm behaves similarly to PrunTree et J4.8 w.r.t. the size of the trees. However, when the size is significantly different different, it can be seen that the accuracy of ΠTree is better. O-ΠTree is less stable and may in three cases induce the largest threes. O-ΠTree is up to 10 times slower than ΠTree when considering all the examples. However, the average update time of the decision tree is negligible (in the worst case it is 100 times faster that ΠTree). It confirms the applicability of O-ΠTree for online learning.

6. Conclusion

In this paper we have proposed an extension of the log-based information gain that takes into account the confidence intervals of the estimates of the frequencies in case of a limited amount of data, thanks to the use of possibility-based representation of the family of probability distribution that agree with the data. This gain function leads us to the learning of well-balanced decision trees, which size are comparable to the ones obtained with a post pruning algorithm. Note that post-pruning algorithm could also benefit from the possibilistic cumulative entropy. It also allows us to propose an incremental version of the algorithm. Experiments show that possibilistic cumulative entropy is a valuable quality measure for decision trees, and that our main algorithm performs very well in comparison with the classical approach. They also confirm the interest of the online algorithm. In the future, we plan to incorporate the treatment of uncertainty around numerical thresholds (like (Katz et al., 2012)) into possibilistic cumulative entropy in order to have a complete handling of uncertainty in the entropy calculus. The approach could also be easily extended to the learning of regression trees, especially for online computation.

References

Abellàn, J. and Moral, S. Upper entropy of credal sets. applications to credal classification. *International Journal of Approximate Reasoning*, 39:235–255, 2005.

Agresti, A. and Coull, B.A. Approximate Is Better than ”Exact” for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126, May 1998.

Bernard, J.M. An introduction to the imprecise dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(23):123 – 150, 2005. Imprecise Probabilities and Their Applications.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. *Classification and Regression Trees*. Chapman & Hall, New York, NY, 1984.

Buntine, W. and Niblett, T. A further comparison of splitting rules for decision-tree induction. *Machine Learning*, 8(1):75–85, 1992.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006.

Domingos, P. and Geoff, H. Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’00, pp. 71–80, New York, NY, USA, 2000. ACM.

Dubois, D. Possibility theory and statistical reasoning. *Computational Statistics and Data Analysis*, 51:47–69, 2006.

Dubois, D. and Hüllermeier, E. Comparing probability measures using possibility theory: A notion of relative peakedness. *International Journal of Approximate Reasoning*, 45(2):364–385, 2007.

Dubois, D. and Prade, H. When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49:65–74, 1992.

Dubois, D., Prade, H., and Sandri, S. On possibility / probability transformations. In Lowen, R. and Roubens, M. (eds.), *Fuzzy Logic - State of the Art*, pp. 103–112. Kluwer Acad. Publ., 1993.

Dubois, D., Foulloy, L., Mauris, G., and Prade, H. Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing*, 10:273–297, 2004.

Esposito, F., Malerba, D., and Semeraro, G. A comparative analysis of methods for pruning decision trees. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5):476–491, May 1997.

Gama, J., Fernandes, R., and Rocha, R. Decision trees for mining data streams. *Intell. Data Anal.*, 10(1):23–45, January 2006.

Grassberger, P. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(67):369 – 373, 1988.

Hüllermeier, E. Possibilistic induction in decision-tree learning. In *Proceedings of the 13th European Conference on Machine Learning*, ECML ’02, pp. 173–184, London, UK, UK, 2002. Springer-Verlag.

Jenhani, I., Ben Amor, N., and Elouedi, Z. Decision trees as possibilistic classifiers. *Inter. J. of Approximate Reasoning*, 48(3):784–807, 2008.

Katz, G., Shabtai, A., Rokach, L., and Ofek, N. Confdtree: Improving decision trees using confidence intervals. In *ICDM*, pp. 339–348, 2012.

Lee, A. Bujaand Y.-S. Data mining criteria for tree-based regression and classification. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’01, New York, NY, USA, 2001. ACM.

Nowozin, S. Improved information gain estimates for decision tree induction. In Langford, John and Pineau, Joelle (eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 297–304. ACM, 2012.

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

Schlimmer, J. C. and Fisher, D.H. A case study of incremental concept induction. In *AAAI*, pp. 496–501, 1986.

Serrurier, M. and Prade, H. An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. *Int. J. Approx. Reasoning*, 54(7):919–933, 2013.

Utgoff, P.E. Incremental induction of decision trees. *Machine Learning*, 4(2):161–186, 1989.

Zadeh, L. A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 1:3–25, 1978.