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

M. Serrurier, H. Prade
IRIT, Toulouse, France

July 7, 2015
Motivations
Possibility distribution as a probability distribution family
Possibilistic loss function and entropy
Application to the learning of decision trees
Conclusions and perspectives
Motivations

- Decision tree learning (as numerous machine learning algorithms) is based on entropy evaluation.

- **Limitations:**
  - The examples become rarer when we go deeper downward the tree, and the faithfulness of entropy decreases
  - Splitting a node always decreases the weighted entropy of the leaves obtained

- Early-stopping and pruning still depend on the entropy previously computed

- **Idea:** describe loss and entropy functions that take into account uncertainty around distribution parameters
Possibility theory

- Possibility distribution $\pi$ is a mapping from $\Omega = \{C_1, \ldots, C_q\}$ to $[0, 1]$.
- Possibility measure: $\forall A \subseteq \Omega, \Pi(A) = \sup_{x \in A} \pi(x)$
  - $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$
  - $\Pi(A \cap B) \leq \min(\Pi(A), \Pi(B))$
- States of knowledge:
  - Complete knowledge: $\exists x \in \Omega$ such as $\pi(x) = 1$ and $\forall y \in \Omega, y \neq x, \pi(y) = 0$
  - Total ignorance: $\forall x \in \Omega, \pi(x) = 1$. 
Possibility distribution and upper bound of probability distribution

- Qualitative interpretation: description of imprecise concept (cheap, young, ...)
- Probabilistic interpretation: upper bound of a probability distribution family:
  \[ \mathcal{P}(\pi) = \{ p \in \mathcal{P}, \forall A \subseteq \Omega, P(A) \leq \Pi(A) \} \].

- States of knowledge:
  - Complete knowledge: no uncertainty.
  - Total ignorance: all probability distributions are possible.
Probability-possibility transformation

- $\sigma$-specificity (discrete case) : $\pi \preceq_\sigma \pi'$, if and only if exists a permutation $\sigma \in S_q$ such as:

$$\pi \preceq_\sigma \pi' \iff \forall x \in \Omega, \pi(x) \leq \pi'(\sigma(x))$$

- Specificity reflects the amount of information encoded by the distribution.

- Probability-possibility transformation ($T_p^*$) : the most $\sigma$-specific possibility distribution which bounds the distribution

- $T_p^*$ is a cumulative function of $p$ where $\sigma^* \in S_q$ follows the probability increasing order
Probability-possibility transformation: example

$p$

$T_p^{\sigma^1}$

$T_p^{\sigma^2}$

$T^*$
Why using possibility distribution?

- Probability-possibility transformation → loss of information but ...
- Possibility distribution can encode different states of knowledge from complete knowledge to total ignorance
- \( T_p^* \) + uncertainty around parameters of \( p \) → less specific possibility distribution

Comparing possibility = comparing entropy

\[ T_p^* \preceq_\sigma T_p'^* \Rightarrow \mathcal{H}(p) \leq \mathcal{H}(p') \]

Goal: describe loss and entropy functions that support the \( \sigma \)-specificity order and the probabilistic interpretation of possibility distribution
Possibilistic cumulative entropy for a limited set of data

- Given $p(c)$ estimated from $n$ pieces of data, we compute the upper bound $p_{\gamma,n}^*$ of the $(1 - \gamma)\%$ confidence interval with Agresti-Coull method.

- Possibility distribution as an upper bound of the frequency distribution

\[ \pi_{\rho,n}(C_j) = P_{\gamma,n}^*(\bigcup_{i=1}^{j} C_{\sigma(i)}) \]

where $\sigma \in S_\rho$ follows the probability order.
Example

- $p(C_1) = 0.5$, $p(C_2) = 0.2$ and $p(C_3) = 0.3$.
- $n = 10$ ($\gamma = 0.05$)
- $\pi_{p,10}^{0.05}(C_1) = P_{0.05,10}^*(C_1 \cup C_2 \cup C_3) = 1$
- $\pi_{p,10}^{0.05}(C_2) = p_{0.05,10}^*(C_2) = 0.52$
- $\pi_{p,10}^{0.05}(C_3) = P_{0.05,10}^*(C_2 \cup C_3) = 0.76$

Properties:

- $p \in \mathcal{P}(\pi_{\gamma}^p, n)$
- $\forall n > 0$, $\pi_p^* \leq \pi_{\gamma}^p$
- $\lim_{n \to \infty} \pi_{\gamma}^p_n = \pi_p^*$
Probabilistic loss functions and entropy

- Loss functions $\mathcal{L}(f, X)$ measure adequateness between data $X = \{x_1, \ldots, x_n\}$ and a distribution $f$.
- Loss function $\mathcal{L}(f, X)$ is linear w.r.t. $X$: $\mathcal{L}(f, X) = \frac{\sum_{i=1}^{n} \mathcal{L}(f, x_i)}{n}$
- Log loss: $\mathcal{L}_{\text{log}}(p|X) = -\sum_{j=1}^{q} \alpha_j \log(p_j)$.
- The entropy is the loss function value of the frequency distribution (i.e. $\mathcal{H}(p^\alpha) = \mathcal{L}(p^\alpha|X)$)
- Entropy is maximal for the uniform distribution
- Entropy is minimal when all the data pertain to the same class
Possibilistic log-loss function

- Principle given a possibility distribution $\pi$:
  - consider $\sigma$ s.a. $\pi(C_{\sigma(1)}) \leq \ldots \leq \pi(C_{\sigma(q)})$
  - consider $BC_j = \bigcup_{i=1}^{j} C_{\sigma(i)}$ and $\overline{BC_j}$ as binary event space
  - $(\pi(C_{\sigma(j)}), 1 - \pi(C_{\sigma(j)}))$ is a probability distribution on $\Omega_j = \{BC_j, \overline{BC_j}\}$
  - apply re-scaled loss function to each Bernoulli distribution

- Poss-log loss:
  $$L_{\pi-l}(\pi|X)) = -\sum_{j=1}^{q} \left( \frac{\text{cdf}_j}{2} \ast \log\left(\frac{\pi_j}{2}\right) + (1 - \frac{\text{cdf}_j}{2}) \ast \log(1 - \frac{\pi_j}{2}) \right)$$

- Possibilistic entropy:
  $$H_{\pi-l}(p, \pi) = -\sum_{j=1}^{q} \frac{T_p^*(C_j)}{2} \ast \log\left(\frac{\pi(C_j)}{2}\right) + (1 - \frac{T_p^*(C_j)}{2}) \ast \log\left(1 - \frac{\pi(C_j)}{2}\right)$$
  $$H^*_{\pi-l}(p, n, \gamma) = H_{\pi-l}(p, \pi^{\gamma}_{p,n})$$
**Possibilistic log-loss function: properties**

**Linearity**

\[ \mathcal{L}_\pi \text{ is linear with respect to } X \]

**Optimality for probability possibility transformation**

we have \(\arg\min(\mathcal{L}_\pi(\pi|X)) = T_{p^\alpha}^*\) (where \(p^\alpha\) is the frequency distribution).

**Increases when uncertainty increases**

given \(n' \leq n\) we have

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

**Specificity order**

\[
\forall \gamma \in ]0, 1[, T_p^* \preceq T_{p'}^* \Rightarrow \mathcal{H}_{\pi-1}^*(p, n, \gamma) \leq \mathcal{H}_{\pi-1}^*(p', n, \gamma)
\]
Algorithm

- Principle: recursively choose the attribute that maximize the gain function
- Log gain function:

\[ G(Z, A) = H(p_Z) - \sum_{k=1}^{r} \frac{|v_k|}{n} H(p_{v_k}) \]

- Possibilistic gain function:

\[ G^\pi_\gamma(Z, A) = H^*_{\pi-1}(p_Z, n, \gamma) - \sum_{k=1}^{r} \frac{|v_k|}{n} H^*_{\pi-1}(p_{v_k}, |v_k|, DS(\gamma, r)) \]
Experiments: Entropy vs size

size vs log loss

size vs possibilistic log-loss
Experiments: entropy vs accuracy

Accuracy vs log-loss

Accuracy vs possibilistic log-loss
Advantages of the approach

• Significant choices of split
• Statistically relevant stopping criterion (negative gain)
• Reasonable estimator of the performances of a decision tree
• Provide well sized and well balanced trees
• Perform statistically better than classical approach and naive pruning
• Statistically equivalent performance (accuracy and size) than J48
Online algorithm

Algorithm:
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.

Advantages:
- Incremental algorithm
- Built new leaves only when the split gives a statistical significant gain
- Only consider the leaf concerned by the new example
Conclusions

- Possibility loss functions and entropies
  - Agree with the probabilistic view of possibility theory
  - Reflect both the entropy of a probability distribution and the uncertainty around the parameters

- Application to decision trees:
  - Provides well balanced and well sized trees
  - Avoids over-fitting
  - Simple and efficient online algorithm

- Perspective:
  - Handle uncertainty around numerical threshold
  - Extension to regression trees