# Cautious Decision-Making for Tree Ensembles

**Haifei Zhang, Benjamin Quost and Marie-Hélène Masson**

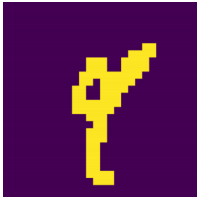
Université de technologie de Compiègne, France
Université de Picardie Jules Verne, France
Laboratoire Heudiasyc, UMR-CNRS 7253


**ECSQARU 2023, 20 September, Arras, France**

# Content

## Why do we need cautious classifiers ?

Cat or dog ?



Number 4 or 9 ?

- deal with uncertainty in the data or the model itself
- reduce the risk of errors in high-stakes applications
- make models robust to minor fluctuations or outliers in the data

## Random forests for cautious decisions

**In random forests** :

- trees provide precise probability distributions
- trees are aggregated by averaging or voting to build a single probability distribution
- precise decisions are made according to the maximum expected utility across all classes

## Random forests for cautious decisions

**In random forests** :

- trees provide precise probability distributions
- trees are aggregated by averaging or voting to build a single probability distribution
- precise decisions are made according to the maximum expected utility across all classes

**In cautious random forests :**

- imprecise trees provide probability intervals based on the imprecise Dirichlet model
- imprecise trees are aggregated by generalized averaging or voting (contribution #1)
- cautious decisions are efficiently made according to the maximum lower expected utility (contribution #2)

# Content

## Cautious classifiers

- Input space : $\mathcal{X}$
- Class labels : $\Omega = \{c_1, \ldots, c_K\}$
- Output space : $\mathcal{Y} = \mathscr{P}(\Omega)$, the power set of $\Omega$
- Cautious classifier $h : \mathcal{X} \to \mathscr{P}(\Omega)$
- cautious prediction for instance $x \in \mathcal{X}$ : $h(x) \subseteq \Omega$

## Evaluation metrics for cautious classifiers

- **determinacy** : the proportion of samples that are determinately classified
- **single-set accuracy** : the proportion of correct determinate decisions
- **set accuracy** : the proportion of indeterminate predictions that contains the actual class
- **set size** : the average size of indeterminate predictions
- **discounted utility** : the expected utility of making a correct decision, discounted by the size of the predicted set, for test instance $(x, y)$ and $h(x) = \hat{Y}$,

$$u_\alpha(\hat{Y}, y) = d_\alpha(|\hat{Y}|)\mathbb{1}(y \in \hat{Y}). \tag{1}$$

Two commonly used discounted utilities are

$$u_{65}(|\hat{Y}|) = \left(\frac{1.6}{|\hat{Y}|} - \frac{0.6}{|\hat{Y}|^2}\right)\mathbb{1}(y \in \hat{Y}) \text{ and } u_{80}(|\hat{Y}|) = \left(\frac{2.2}{|\hat{Y}|} - \frac{1.2}{|\hat{Y}|^2}\right)\mathbb{1}(y \in \hat{Y}).$$

# Example of utility matrix

Table – Example of different discounted utility matrices

| $\hat{Y}$ | $u_{65}(\lvert\hat{Y}\rvert)$ | | | $u_{80}(\lvert\hat{Y}\rvert)$ | | |
|---|---|---|---|---|---|---|
| | $c_1$ | $c_2$ | $c_3$ | $c_1$ | $c_2$ | $c_3$ |
| $\{c_1\}$ | 1 | 0 | 0 | 1 | 0 | 0 |
| $\{c_2\}$ | 0 | 1 | 0 | 0 | 1 | 0 |
| $\{c_3\}$ | 0 | 0 | 1 | 0 | 0 | 1 |
| $\{c_1, c_2\}$ | 0.65 | 0.65 | 0 | 0.8 | 0.8 | 0 |
| $\{c_1, c_3\}$ | 0.65 | 0 | 0.65 | 0.8 | 0 | 0.8 |
| $\{c_2, c_3\}$ | 0 | 0.65 | 0.65 | 0 | 0.8 | 0.8 |
| $\{c_1, c_2, c_3\}$ | 0.467 | 0.467 | 0.467 | 0.6 | 0.6 | 0.6 |

## Theory of belief functions

- $\Omega = \{c_1, \ldots, c_K\}$ : the frame of discernment
- $m : 2^\Omega \to [0, 1]$ : a mass function such that $\sum_{A \subseteq \Omega} m(A) = 1$

For any $A \subseteq \Omega$, its belief and plausibility degrees are

$$Bel(A) = \sum_{B \subseteq A} m(B), \tag{2}$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \tag{3}$$

## Decision-making with belief functions

Given a test instance $x$, set of class $\Omega$, and an associated mass function $m$,

the maximum lower expected utility criterion makes the decision as

$$\hat{Y} = \arg\max_{A \subseteq \Omega} \underline{EU}(m, A, U) = \arg\max_{A \subseteq \Omega} \sum_{B \subseteq \Omega} m(B) \min_{c_j \in B} u_{Aj}, \tag{4}$$

where $u_{Aj}$ is the utility when $A$ is taken as a prediction and $c_j$ is the actual class for $x$.

## Content

## Content

## Imprecise trees induced by IDM

- $\Omega = \{c_1, \ldots, c_K\}$ : class labels
- $H = \{h_1, \ldots, h_T\}$ : a random forest trained for multi-class classification on $\Omega$
- $n_{tj}(x)$ : the number of training samples of class $c_j$ in the leaf of $h_t$ where $x$ falls into
- $N_t = \sum_{j=1}^{K} n_{tj}(x)$ : the total number of training samples in the leaf of $h_t$ where $x$ falls into

The probability intervals for $\mathbb{P}(c_j|x, h_t)$ by applying the imprecise Dirichlet model is :

$$\mathcal{I}_{tj} = \left[\underline{p}_{tj}, \overline{p}_{tj}\right] = \left[\frac{n_{tj}}{N_t + s}, \frac{n_{tj} + s}{N_t + s}\right], \; j = 1, \ldots, K, \tag{5}$$

where $s$ is interpreted as the number of virtual samples with unknown actual classes.

## How to combine imprecise trees ?

The IDM probability intervals $\mathcal{I}_{tj} = \left[\underline{p}_{tj}, \overline{p}_{tj}\right]$ provided by each tree can be seen as a quasi-Bayesian mass function :

$$m_t(\{c_j\}) = \underline{p}_{tj},\ j = 1, \ldots, K, \quad m_t(\Omega) = 1 - \sum_{j=1}^{K} m_t(\{c_j\}). \tag{6}$$

**How to aggregate them into a single mass function ?**

## Generalization of averaging

**From averaging probability distributions to averaging mass functions :**

$$m(\{c_j\}) = \frac{\sum_{t=1}^{T} m_t(\{c_j\})}{T},\ j = 1, \ldots, K \quad m(\Omega) = \frac{\sum_{t=1}^{T} m_t(\Omega)}{T}. \tag{7}$$

**The mass function $m$ is still quasi-Bayesian.**

## Generalization of voting

**From voting for a single class to voting for a subset of classes :**

$$m(A) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(ID(m_t) = A), \tag{8}$$

where $ID(\cdot)$ is the interval dominance criterion that returns the set of non-dominated classes for each tree.

**The mass function $m$ is no longer quasi-Bayesian.**

## Content

## Efficient maximization of the lower expected utility

- $x$ : test instance
- $m$ : mass function provided by the aggregation of imprecise trees
- $u_{Aj} = d_\alpha(|A|)\mathbb{1}(c_j \in A)$ : utility of taking $A \subseteq \Omega$ as cautious prediction while $c_j \in \Omega$ is the real label

The maximizer of the lower expected utility in Eq. (4) can be written as

$$\hat{Y} = \arg\max_{A \subseteq \Omega} \sum_{B \subseteq \Omega} m(B) \min_{c_j \in B} u_{Aj} = \arg\max_{A \subseteq \Omega} d_\alpha(|A|) Bel(A). \tag{9}$$

**Search for the maximizer of the lower expected utility by scanning subsets of classes with increasing cardinality.**

## Cautious decision-making with the averaged mass function (CDM_Ave)

**Decision-making process :**

search for the maximizer of $\underline{EU}$ by scanning subsets of classes with increasing cardinality

1. sort classes by decreasing mass : $m(\{c_{(j)}\}) \geq m(\{c_{(j+1)}\})$, for $j = 1, \ldots, K-1$
2. add classes one-by-one to the candidate prediction and calculate the lower expected utility
3. keep the prediction with the highest lower expected utility

**Remarks :**

the averaged mass function is still quasi-Bayesian,
thus, the above decision-making process can be solved in complexity $O(|\Omega|)$.

## Cautious decision-making with the averaged mass function (CDM_Ave)

### Example

Assume the averaged mass function on $\Omega = \{c_1, c_2, c_3, c_4\}$ is given as follows :

$$m(\{c_1\}) = 0.32,\ m(\{c_2\}) = 0.48,\ m(\{c_3\}) = 0.04,\ m(\{c_4\}) = 0.06,\ m(\Omega) = 0.05.$$

The classes ordered by decreasing mass are thus $\{c_2, c_1, c_4, c_3\}$. These classes are added to the prediction one by one and the corresponding expected lower discounted utilities (using $d_{u_{65}}$) are calculated :

- $\underline{EU}(\{c_2\}) = d_{u_{65}}(1)Bel(\{c_2\}) = 1 \times 0.48 = 0.48$
- $\underline{EU}(\{c_2, c_1\}) = d_{u_{65}}(2)Bel(\{c_2, c_1\}) = 0.65 \times 0.8 = 0.52$
- $\underline{EU}(\{c_2, c_1, c_4\}) = d_{u_{65}}(3)Bel(\{c_2, c_1, c_4\}) = 0.4667 \times 0.86 = 0.401$
- $\underline{EU}(\{c_2, c_1, c_4, c_3\}) = d_{u_{65}}(4)Bel(\Omega) = 0.3625 \times 1 = 0.3625.$

We can find that $\{c_2, c_1\}$ reaches the maximum expected lower discounted utility : thus, the cautious prediction made is $\hat{Y} = \{c_2, c_1\}$.

## Cautious decision-making with the mass function of generalized voting (CDM_Vote)

**Decision-making process :**

search for the maximizer of $\underline{EU}$ by scanning subsets of classes with increasing cardinality.

**Remarks :**

- the mass function obtained by the generalized voting is no longer quasi-Bayesian
- maximizing the lower expected utility requires in principle checking all subsets of $\Omega$
- it is intractable for data with a large number of classes

## Cautious decision-making with the mass function of generalized voting (CDM_Vote)

**To reduce the complexity, we introduce three tricks :**

1. restrict the decision to subsets $A \subseteq \Omega$ with cardinality $|A| \leq M \leq K$

## Cautious decision-making with the mass function of generalized voting (CDM_Vote)

**To reduce the complexity, we introduce three tricks :**

1. restrict the decision to subsets $A \subseteq \Omega$ with cardinality $|A| \leq M \leq K$
2. stop the procedure when larger subsets are known not further to improve the lower expected utility

## Cautious decision-making with the mass function of generalized voting (CDM_Vote)

**To reduce the complexity, we introduce three tricks :**

1. restrict the decision to subsets $A \subseteq \Omega$ with cardinality $|A| \leq M \leq K$
2. stop the procedure when larger subsets are known not further to improve the lower expected utility
3. for a given cardinality $i$, only subsets $A$ composed of classes appearing in focal elements $B$ such that $|B| \leq i$ need to be considered

## Cautious decision-making with the mass function of generalized voting (CDM_Vote)

### Example

Assume the mass function on $\Omega = \{c_1, c_2, c_3, c_4\}$ obtained via Eq. (8) is :

$$m(\{c_1\}) = 0.15,\ m(\{c_2\}) = 0.25,$$
$$m(\{c_1, c_2\}) = 0.35,\ m(\{c_1, c_3\}) = 0.05,\ m(\{c_2, c_3\}) = 0.1,$$
$$m(\{c_2, c_3, c_4\}) = 0.05,\ m(\Omega) = 0.05.$$

The cautious decision-making process (using $d_{u_{65}}$) would be :

- $i = 1,\ \Omega_1 = \{c_1, c_2\}$ :

$$\underline{EU}(\{c_1\}) = d_{u_{65}}(1) Bel(\{c_1\}) = 1 \times 0.15 = 0.15,$$
$$\underline{EU}(\{c_2\}) = d_{u_{65}}(1) Bel(\{c_2\}) = 1 \times 0.25 = 0.25,$$
$$A = \{c_2\},\ \underline{EU}(A) = 0.25 < d_{u_{65}}(2) = 0.65,\ \text{continue}$$

## Cautious decision-making with the mass function of generalized voting (CDM_Vote)

### Example (continued)

- $i = 2,\ \Omega_2 = \{c_1, c_2, c_3\}$ :

$$\underline{EU}(\{c_1, c_2\}) = d_{u_{65}}(2)Bel(\{c_1, c_2\}) = 0.65 \times 0.75 = 0.4875,$$
$$\underline{EU}(\{c_1, c_3\}) = d_{u_{65}}(2)Bel(\{c_1, c_3\}) = 0.65 \times 0.2 = 0.13,$$
$$\underline{EU}(\{c_2, c_3\}) = d_{u_{65}}(2)Bel(\{c_2, c_3\}) = 0.65 \times 0.35 = 0.2275,$$
$$A = \{c_1, c_2\},\ \underline{EU}(A) = 0.4875 > d_{u_{65}}(3) = 0.4667,\ \text{stop.}$$

The final set-valued prediction is then $\hat{Y} = \{c_1, c_2\}$ since any subset $B \subseteq \Omega$ with $|B| > 2$ has a smaller lower expected utility than $\hat{Y}$.
Here, class $c_4$ has never been considered because it first appears in the focal element $\{c_2, c_3, c_4\}$ with a cardinality of 3, which was known not to ameliorate the lower expected utility.

# Content

# Decision-making efficiency of CDM_Vote

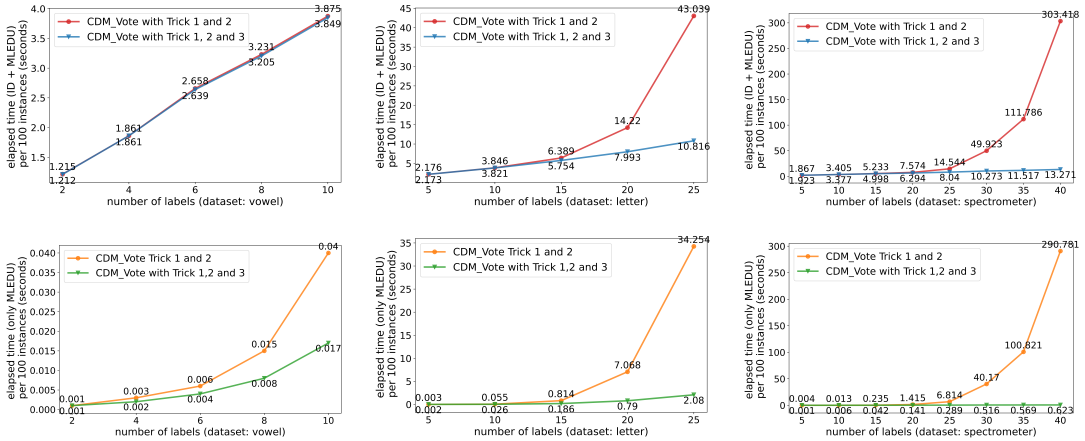



Figure – Decision-making time complexity of CDM_Vote according to the number of labels (for 100 samples). The first row : ID+MLEDU the second row : MLEDU only.

## Performance comparison of different cautious classifiers

- MVA : each tree votes for dominated classes and the classes with minimal opposition are retained
- AVE : probability intervals are averaged across trees and predictions are made by applying the interval dominance approach.

Table – Dataset : vowel (11 labels)

| Criteria | MVA | AVE | CDM_Vote | CDM_AVE |
|---|---|---|---|---|
| Determinacy | **0.995±0.007** | 0.918±0.032 | 0.874±0.036 | 0.867±0.038 |
| Single-set accuracy | 0.952±0.024 | 0.982±0.015 | 0.991±0.013 | **0.994±0.011** |
| Set accuracy | 0.944±0.168 | **0.974±0.063** | 0.967±0.056 | 0.962±0.053 |
| Set size | **2.0±0.0** | 2.418±0.275 | 2.054±0.064 | 2.056±0.064 |
| $u_{65}$ score | **0.950±0.025** | 0.948±0.019 | 0.944±0.016 | 0.941±0.017 |
| $u_{80}$ score | 0.950±0.024 | 0.960±0.017 | **0.963±0.013** | 0.960±0.013 |

## Performance comparison of different cautious classifiers

Table – Dataset : letter (26 labels)

| Criteria | MVA | AVE | CDM_Vote | CDM_AVE |
|---|---|---|---|---|
| Determinacy | **0.988±0.008** | 0.772±0.026 | 0.816±0.026 | 0.811±0.026 |
| Single-set accuracy | 0.861±0.026 | **0.964±0.016** | 0.943±0.018 | 0.949±0.016 |
| Set accuracy | 0.717±0.259 | **0.949±0.030** | 0.710±0.078 | 0.728±0.071 |
| Set size | **2.077±0.208** | 12.197±1.390 | 2.139±0.058 | 2.163±0.062 |
| $u_{65}$ score | 0.855±0.026 | 0.809±0.023 | 0.852±0.021 | **0.856±0.020** |
| $u_{80}$ score | 0.856±0.026 | 0.826±0.022 | 0.871±0.020 | **0.876±0.019** |

## Performance comparison of different cautious classifiers

Table – Dataset : spectrometer (48 labels)

| Criteria | MVA | AVE | CDM_Vote | CDM_AVE |
|---|---|---|---|---|
| Determinacy | **0.978±0.023** | 0.544±0.071 | 0.480±0.063 | 0.499±0.064 |
| Single-set accuracy | 0.550±0.068 | 0.694±0.074 | **0.700±0.076** | 0.690±0.077 |
| Set accuracy | 0.741±0.280 | **0.817±0.080** | 0.722±0.097 | 0.712±0.099 |
| Set size | **2.067±0.222** | 9.582±3.213 | 2.132±0.072 | 2.121±0.065 |
| $u_{65}$ score | 0.545±0.066 | 0.538±0.050 | **0.571±0.051** | 0.568±0.052 |
| $u_{80}$ score | 0.546±0.066 | 0.580±0.052 | **0.626±0.055** | 0.621±0.055 |

# Content

## Conclusion

In this work, we

- generalized averaging and voting to trees providing probability intervals
- proposed a cautious decision-making strategy by maximizing the lower expected utility
- proposed an efficient implementation for handling large numbers of classes

## perspectives

In future works, we will

- investigate how to make CDM_Vote more efficient and tractable
- explore other less conservative strategies, e.g., the Hurwitz criterion
- compare our cautious decision-making strategies with other cautious classifiers beyond tree-based models.

# Thanks !