Cautious Random Forests: a new Decision Strategy and some Experiments

Haifei Zhang

Benjamin Quost, Marie-Hélène Masson

09 July 2021, ISIPTA21 Granada Spain

Context

- Random forest : an accurate classification strategy
- Available information is scarce
- Large conflict between the decision tree outputs

Purpose : provide cautious (set-valued) predictions to guarantee robustness

Methods applied

- Imprecise Dirichlet Model (IDM)
- Belief Functions Theory

Information about a leaf

1. A forest $F = \{C_1, \ldots, C_t, \ldots, C_T\}$ has been trained for a binary classification problem
2. A test instance x_i is associated with a set of leaves, noted as $R = \{L_i^1, \ldots, L_i^t, \ldots, L_i^T\}$
3. In each leaf, n_i^t and N_i^t are the number of samples of category 1 and the total number of training samples

IDM intervals

The interval-valued estimation $I_i^t = [\underline{p}_{i,1}^t, \overline{p}_{i,1}^t]$ of $p_{i,1}^t = \mathbb{P}(y_i = 1 | x_i, t)$ is :

$$I_i^t = \left[\frac{n_i^t}{N_i^t + s}, \frac{n_i^t + s}{N_i^t + s} \right] \quad t = 1, \ldots, T$$

Tree Aggregation

For an instance x_i, each interval I_i^t provided by a tree is regarded as a focal element associated with a mass m_i^t on the interval $[0, 1]$. The belief and plausibility of the event $p_{i,1} \geq 0.5$ can be defined as :

$$bel_{i,1} = bel(p_{i,1} \in [0.5, 1]) = \sum_{t=1}^{T} m_i^t \mathbb{I}(\underline{p}_{i,1}^t \geq 0.5),$$

$$pl_{i,1} = pl(p_{i,1} \in]0.5, 1]) = \sum_{t=1}^{T} m_i^t \mathbb{I}(\overline{p}_{i,1}^t > 0.5),$$

Decision Strategy : interval dominance

$$\hat{y}_i = \begin{cases} 0, \text{if } pl_{i,1} < 0.5 \\ 1, \text{if } bel_{i,1} \geq 0.5 \\ \{0, 1\}, \text{otherwise} \end{cases}$$

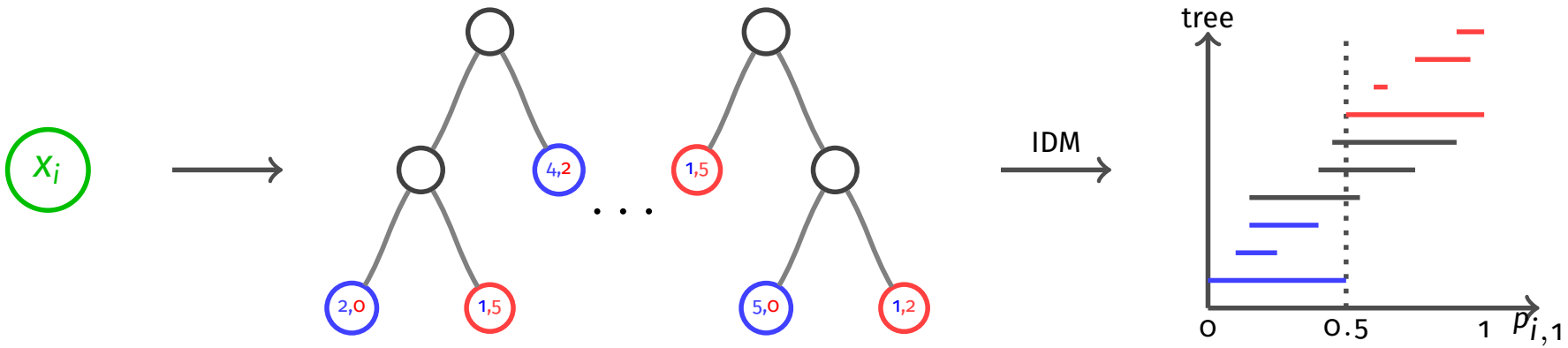
Mass assignment strategies

1. Equal mass : $m_i^t = \frac{1}{T}$, $\forall t = 1, \ldots, T$, where T is the number of trees in the forest

2. Leaf size based mass : $m_i^t = \frac{N_i^t}{\sum_{j=1}^{T} N_i^j}$, $\forall t = 1, \ldots, T$, where N_i^t is the number of training samples in the same leaf as x_i for tree t

3. Interval uncertainty based mass : $m_i^t = \frac{1-u_i^t}{\sum_{j=1}^{T}(1-u_i^j)}$, with $u_i^t = \frac{s}{N_i^t+s}$, $\forall t = 1, \ldots, T$, the level of epistemic uncertainty for instance x_i and for tree t

Inference process summary



(Input) (Random Forest) (Probability Intervals)

Aggregation

$bel_{i,1}$ $pl_{i,1}$

e.g.,0.4 e.g.,0.7

Interval Dominance

$$\hat{y}_i = \begin{cases} 0, \text{ if } pl_{i,1} < 0.5 \\ 1, \text{ if } bel_{i,1} \geq 0.5 \\ \{0, 1\}, \text{ otherwise} \end{cases}$$

(Belief and Plausibility) (Prediction)

Experimentation :

- Baseline : tree aggregation consists in averaging the lower and upper probability bounds over all trees : $bel_{i,1} = ave(\underline{p}^t_{i,1})$ and $pl_{i,1} = ave(\overline{p}^t_{i,1})$
- Data sets : eight common binary classification data sets from UCI
- Comparison : U65 score with different value of s

Results

1. Our imprecise classifier tends to be less cautious than baseline and provides a better compromise (U65) when s is large
2. Our model is robust to high epistemic uncertainty
3. Computing the mass based on epistemic uncertainty proves to be much more fruitful when s is large

Conclusion and perspectives

A cautious random forest model is proposed for binary imprecise classification

1. A new mechanism based on belief functions theory to aggregate IDM probability intervals of trees
2. Different methods for mass assignment, e.g., equal mass, leaf size mass and interval uncertainty mass

Perspectives :

- Other methods for mass assignment : automatically learn tree weights from data
- Extension to cautious multi-class classification
- Explanations for indeterminate decisions

Thanks for your attention!