A-PETE: Adaptive Prototype Explanations of Tree Ensembles

Jacek Karolczak, Jerzy Stefanowski
Poznan University of Technology
jacek.karolczak@student.put.poznan.pl, jerzy.stefanowski@cs.put.poznan.pl

Introduction

This study aims explaining opaque ensembles of tree classifiers models. The need of explanations is addressed through prototypes – representative instances that illuminate model behaviour. Prototypes offer both global insights into model behaviour and local explanations for individual decisions [1].

Distance for tree ensembles

Let t represent the number of trees in the tree ensemble (TE). The i-th tree ($i \in [t]$) partitions the feature space into regions $R_{i,j}$, each corresponding to a leaf $\tau_{i,j}$. Each tree induces an individual classifier assigning each point $x \in X$ to a single region $R_{i,j}$ [3]:

$$c_i^{\text{Tree}}(x) = \sum_{j=1}^{\tau_i} \alpha_{i,j} \mathbb{1}(x \in R_{i,j}) \,,$$

where $\alpha_{i,j}$ is the predicted value in the j-th leaf of the i-th tree. $\mathbb{1}$ denotes the indicator function. The tree ensemble classifier is the average over all trees:

$$c^{\text{TE}}(x) = \frac{1}{t} \sum_{i=1}^{t} c_i^{\text{Tree}} = \frac{1}{t} \sum_{i=1}^{t} \sum_{j=1}^{\tau_i} \alpha_{i,j} \mathbb{1}(x \in R_{i,j})$$

Thus, the proximity of two instances x_1 and x_2 is given as the mean number of trees in which both instances land in the same leaf and can be expressed as

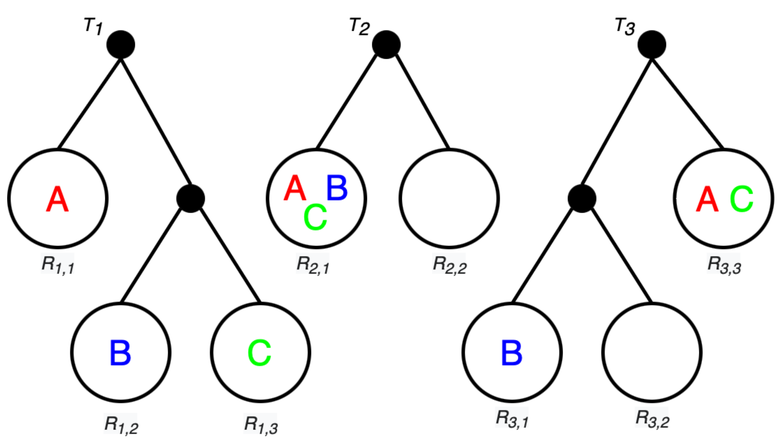
$$p^{\text{TE}}(x_1, x_2) = \frac{1}{t} \sum_{i=1}^{t} \sum_{j=1}^{\tau_i} \mathbb{1}(x_1 \in R_{i,j}) \mathbb{1}(x_2 \in R_{i,j})$$

The distance metric can be derived from the proximity function:

$$d^{\text{TE}}(x_1, x_2) = 1 - p^{\text{TE}}(x_1, x_2)$$

Toy example

Figure 1: An artificial example showing how proximity is computed with respect to a tree ensemble.



$p^{\text{TE}}(\mathbf{A}, \mathbf{B}) = \frac{1}{3}$, $p^{\text{TE}}(\mathbf{A}, \mathbf{C}) = \frac{2}{3}$, $p^{\text{TE}}(\mathbf{B}, \mathbf{C}) = \frac{1}{3}$

A-PETE

We propose the Adaptive Prototype Explanations of Tree Ensembles (A-PETE), which automatically selects k prototypes. We adopt the greedy submodular prototype selection algorithm (SM-A), which minimises the function:

$$f(P) = \sum_{c}^{C} \sum_{i}^{|X^c|} \min_{p^c \in P^c} d^{\text{TE}}(x_i^c, p^c) \,.$$

The main novelty is that our algorithm maintains the difference Δ between consecutive objective function changes to automise prototype selection process.

Algorithm 1: Adaptive Prototype Explanations of Tree Ensembles (A-PETE).

Input : Set of points X, distance function $d : X^2 \mapsto [0, 1]$, class assignment $c : X \mapsto [q]$, control parameter $\alpha \in (0, 1)$
Output: Set of prototypes P

Create set of phantom exemplars $P' = \{p'_1, ..., p'_q\}$ and set $d(p'_i, x) = d(x, p'_i) = 1$ for all $x \in X$
$\Delta \leftarrow 0$
P $\leftarrow \varnothing$
while *True* **do**
  $x^* \leftarrow \arg\max_{x \in X} [f(P') - f(P' \cup P \cup \{x\})]$
  $\Delta' \leftarrow f(P' \cup P) - f(P' \cup P \cup \{x^*\})$
  $P \leftarrow P \cup \{x^*\}$
  if $\frac{|\Delta - \Delta'|}{\Delta'} < \alpha$ **then**
    break
  end
  $\Delta' \leftarrow \Delta$
end

Experimental evaluation

Table 1: The best weighted accuracy [2] achieved using Random Forest (RF) and 1-NN run on prototypes selected using k-means using euclidean distance (K-Means), k-means using distance for tree ensemble (RF-KM), adaptive greedy submodular prototype selection (SM-A), weighted adaptive greedy submodular prototype selection (SM-WA), and Adaptive Prototype Explanations of Tree Ensembles (A-PETE). The number of prototypes in parentheses.

	Breastcancer	Diabetes	Compass	RHC	Mnist	Caltech256
RF	*0.93*	*0.73*	*0.66*	*0.75*	*0.99*	*0.69*
K-Means	0.95 (8)	0.66 (6)	0.63 (10)	0.48 (10)	0.87 (14)	0.58 (6)
RF-KM	0.95 (6)	0.72 (4)	0.28 (16)	0.43 (18)	0.97 (14)	0.70 (10)
SM-A	0.92 (8)	0.74 (3)	0.30 (20)	0.74 (12)	0.97 (14)	0.70 (16)
SM-WA	0.92 (8)	0.72 (2)	0.30 (20)	0.40 (10)	0.97 (11)	0.72 (5)
A-PETE	0.92 (7)	0.73 (5)	0.32 (23)	0.73 (9)	0.97 (19)	0.70 (6)

- A-PETE automise the selection of prototypes.
- The number of yielded prototypes consistently approached the number of prototypes from SM-A [3].
- The accuracy of predictions made by A-PETE is comparable to that of Random Forest, indicating that the prototypes selected by A-PETE effectively capture the crucial information needed to replicate the decision-making process of Random Forest.

References

[1] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and Survey of Explanation Methods for Black Box Models. *Data Mining and Knowledge Discovery*, pages 1–60, 2023.

[2] Dariusz Brzezinski, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczech. Visual-based analysis of classification measures and their properties for class imbalanced problems. *Information Sciences*, 462:242–261, 2018.

[3] Sarah Tan, Matvey Soloviev, Giles Hooker, and Martin T. Wells. Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS '20, page 23–34. Association for Computing Machinery, 2020.