TED2021-129438B-I00

Representative measure approach to assess decision trees reliability

Víctor Toscano Durán [1] Javier Perera Lago [1] Eduardo Paluzo Hidalgo [1,2] Rocío González Díaz [1]

[1]University of Seville {vtoscano,jperera,rogodi}@us.es [2]Universidad Loyola Andalucia epaluzo@{uloyola,us}.es

Objective

Machine learning algorithms are essential in modern data-informed artificial intelligence architecture. Representative datasets are crucial in guiding AI development. Proper training with these datasets reduces model complexity and power consumption while minimizing uncertainties.

This poster employs the ε-**representativeness** measure based on computational topology proposed in [2] to quantify the **similarity** between datasets and its impact on **binary decision tree**. Theoretical results confirm prediction similarities with ε-representativeness, and experiments show a significant correlation with **feature importance rankings**, demonstrating its efficacy for reliable decision trees.

Binary decision trees

A **binary decision tree** is composed of nodes and branches.

- **Node**: Represents a decision based on a feature. A node can be internal, if it is connected to two children nodes, or terminal, if it doesn't have children, representing the endpoints of a branch in the tree. The decision at each node is based on a threshold value.
- **Branch**: Represent the various options or courses of action.

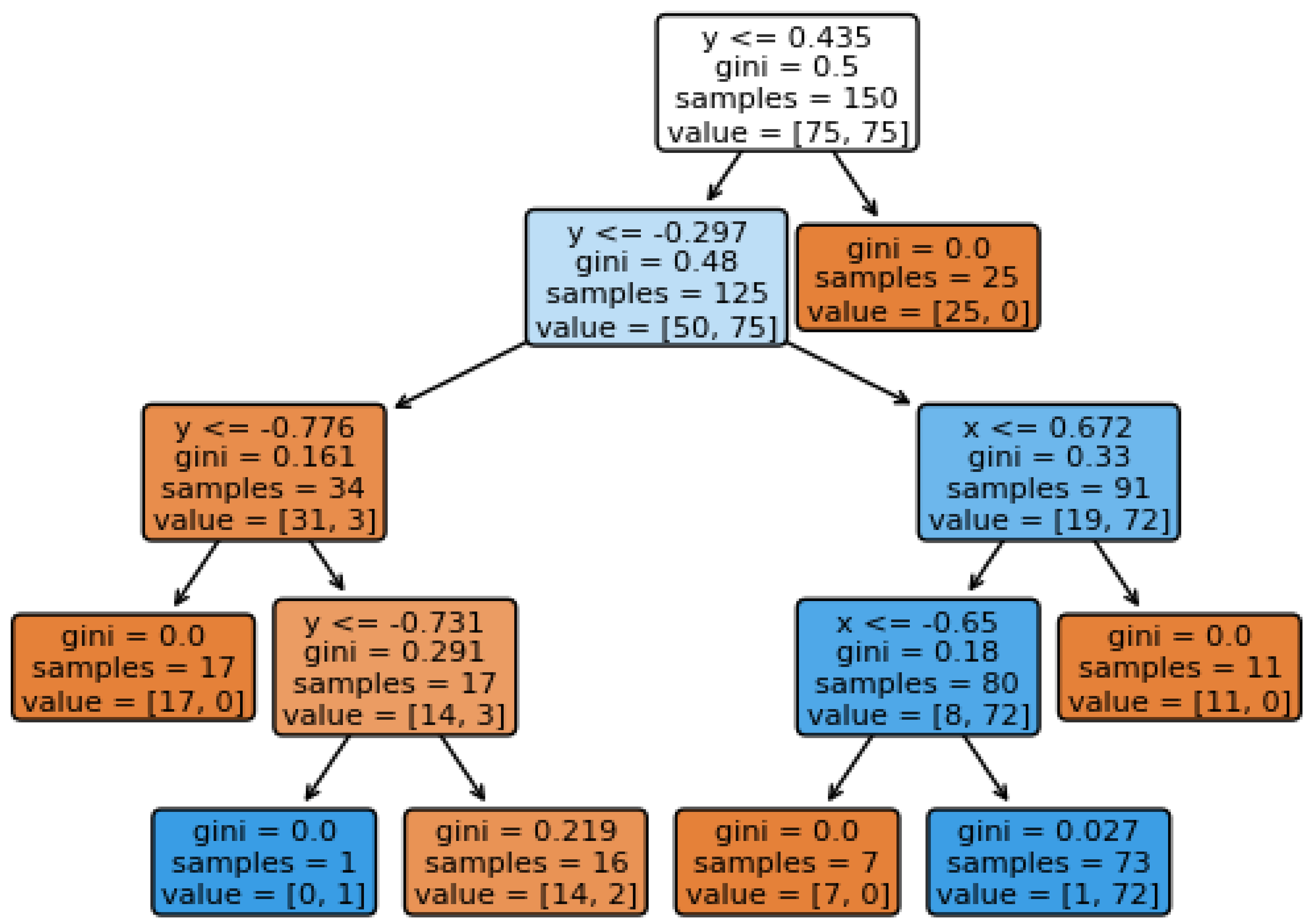

Figure 1. Binary decision tree example. A sample with $x = 0.2$ and $y = 0.1$ will be classify as 1.

The feature and splitting condition are chosen to minimize node **impurity**, a measure of class label homogeneity. The process iterates recursively until a desired depth or no improvement in impurity is possible. We used the Gini index. Let us denote by N_i the number of examples from X reaching the node n_i. The number of class k examples reaching n_i will be denoted by $N_{i,k}$. Then, let us use the following notation: $p_i = N_i/N$, $p_{i,k} = N_{i,k}/N_i$. The Gini index of n_i is $G(n_i) = \sum_{k=1}^{c} p_{i,k}(1-p_{i,k})$. Assume that we fix an impurity measure and we denote it as I. The information gain for an internal node n_i whose two children nodes are n_{i_1} and n_{i_2} is:

$$IG(n_i) = I(n_i) - \frac{N_{i_1}}{N_i} I(n_{i_1}) - \frac{N_{i_2}}{N_i} I(n_{i_2}) \tag{1}$$

Theorem 1

Let $T \in \mathcal{T}$ be a binary decision tree (DT), (X, λ_X) a dataset, and $(\tilde{X}, \lambda_{\tilde{X}})$ a γ-balanced ε-representative dataset of (X, λ_X). If $\varepsilon < M = \min_{i \in I} \mu_i$, then

$$\mathrm{Acc}(T, (X, \lambda_X)) = \mathrm{Acc}(T, (\tilde{X}, \lambda_{\tilde{X}})) \tag{2}$$

Proof. Given $x = (x_1, \cdots, x_n)^T \in X$ and $\tilde{x} = (\tilde{x}_1, \cdots, \tilde{x}_n)^T \in \tilde{X}$ as an ε-representative of x, where $|\tilde{x}_j - x_j| \leq \varepsilon \ \forall j \in \{1, \cdots, d\}$, and assuming $\tilde{x}$ accesses the tree through the root node n_1 and is sent to its left child, we find $0 < t_1 - \tilde{x}_{j_1}$. By definition of margins, $\mu_1 \leq t_1 - \tilde{x}_{j_1}$. Since $\tilde{x}$ is ε-representative of x, $x_{j_1} \leq \tilde{x}_{j_1} + \varepsilon$. Adding these inequalities, $\mu_1 - \varepsilon \leq t_1 - x_{j_1}$. Since $\varepsilon < M \leq \mu_1$, then $0 < t_1 - x_{j_1}$, meaning x also goes left. Similarly, if $\tilde{x}$ goes right, x does too. Following this reasoning, x and $\tilde{x}$ reach the same terminal node n_ℓ and label k_ℓ.

Due to γ-balance, each $\tilde{x} \in \tilde{X}$ of class k represents γ examples from X of class k. All reach the same node in T. Additionally, $N = \gamma \cdot \tilde{N}$, $N_j = \gamma \cdot \tilde{N}_j$, and $N_{j,k} = \gamma \cdot \tilde{N}_{j,k}$. So, $p_j = \tilde{p}_j$ and $p_{j,k} = \tilde{p}_{j,k}$. Hence:

$$\mathrm{Acc}(T, (X, \lambda_X)) = \sum_{j \in L} p_j \cdot p_{j,k_j} = \sum_{j \in L} \tilde{p}_j \cdot \tilde{p}_{j,k_j} = \mathrm{Acc}(T, (\tilde{X}, \lambda_{\tilde{X}})) \tag{3}$$

Acknowledgements

This work was supported in part by the European Union HORIZON-CL4-2021-HUMAN-01-01 under grant agreement 101070028 (REXASI-PRO) and by TED2021-129438B-I00 / AEI/10.13039/501100011033 / Unión Europea NextGenerationEU/PRTR.

ε-representativeness

Given a dataset (X, λ_X) and another dataset $(\tilde{X}, \lambda_{\tilde{X}})$ with the cardinality of $\tilde{X}$ smaller than the one of X, we say that $\tilde{x} \in \tilde{X}$ is an ε-representative of $x \in X$ if $||\tilde{x} - x||_\infty \leq \varepsilon$ and $\lambda_X(x) = \lambda_{\tilde{X}}(\tilde{x})$, and we say that $(\tilde{X}, \lambda_{\tilde{X}})$ is an ε-representative dataset of (X, λ_X) if for all $x \in X$ there exists $\tilde{x} \in \tilde{X}$ that is an ε-representative of x.

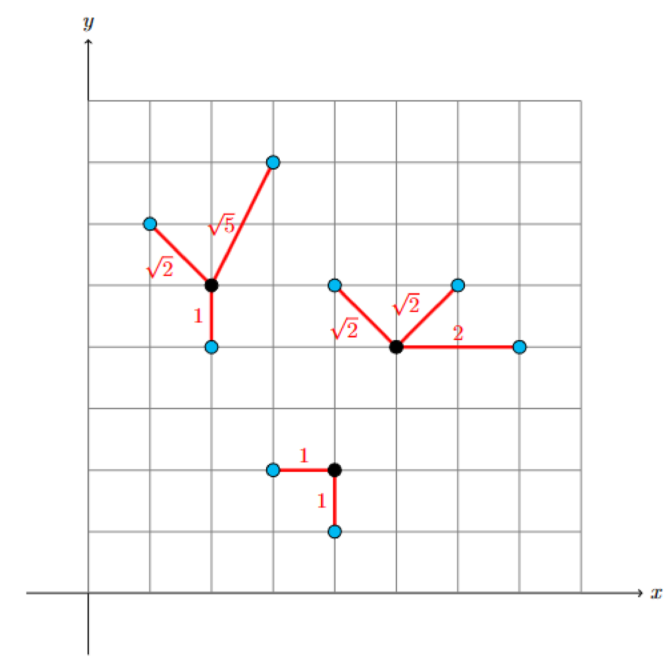

Figure 2. Calculation example of the ε-representativeness of a reduced dataset in relation to a larger one. It involves finding the closest representative point (black) for each original point (blue) and computing the distance. The maximum of these minimum distances, illustrated in red in the graphic, gives the ε value, which is $\sqrt{5}$ in this example.

A dataset $(\tilde{X}, \lambda_{\tilde{X}})$ that is representative of (X, λ_X) is said to be γ-balanced if each $\tilde{x} \in \tilde{X}$ is representative of exactly γ data examples of X and each $x \in X$ is represented by a single example $\tilde{x} \in \tilde{X}$.

Feature importance ordering

Feature importance (FI) quantifies the impact of a particular feature $j \in \{1, \cdots, d\}$ in decreasing the impurity of the decision tree. It is calculated as:

$$FI(j) = \sum_{\substack{i \in I \\ j_i = j}} N_i \cdot IG(n_i) \tag{4}$$

To compare the similarities between binary decision trees, the ordering of feature importance is evaluated using a metric from [1, Section 4.2]. The mean of the absolute differences in feature positions between two ordered sets is calculated:

$$\mathrm{Sim}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left| \mathrm{pos}_x(f_i) - \mathrm{pos}_y(f_i) \right| \tag{5}$$

Results

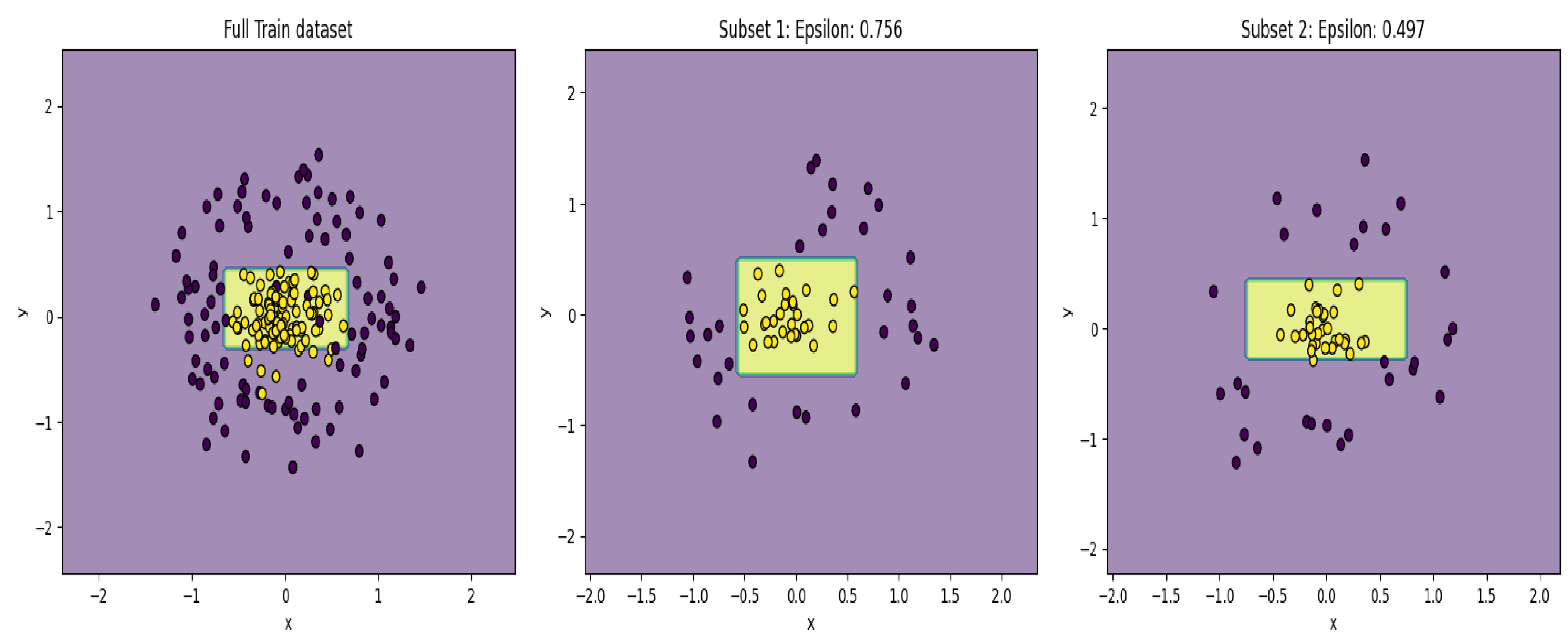

Figure 3. Decision boundaries of binary decision trees (DTs) trained on full training synthetic set and two random subsets. From left to right: (1) the training set; (2) a subset with 40% of the training set and $\varepsilon = 0.756$; (3) a subset with 40% of the training set and $\varepsilon = 0.497$.

The experiment was repeated on 100 subsets of a real dataset. Spearman's correlation(SP) between ε-representativeness and the feature importance metric was calculated, yielding a significant correlation ($Sp = 0.51$, p-value$= 5.2 \times 10^{-8}$).

References

[1] Barrera-Vicent, A., Paluzo-Hidalgo, E., Gutiérrez-Naranjo, M.A.: The metric-aware kernel-width choice for lime. In: Joint Proceedings of the xAI-2023 Late-breaking Work, Demos and Doctoral Consortium co-located with the 1st World Conference on eXplainable Artificial Intelligence (xAI-2023), Lisbon, Portugal, July 26-28, 2023. CEUR Workshop Proceedings, vol. 3554, pp. 117–122 (2023)

[2] Gonzalez-Diaz, R., Gutiérrez-Naranjo, M.A., Paluzo-Hidalgo, E.: Topology-based representative datasets to reduce neural network training resources. Neural Computing and Applications (2022)

[3] Perera-Lago, J., Toscano-Durán, V., Paluzo-Hidalgo, E., Narteni, S., Rucco, M.: Application of the representative measure approach to assess the reliability of decision trees in dealing with unseen vehicle collision data (2024)