Application of the representative measure approach to assess the reliability of decision trees in dealing with unseen vehicle collision data

Víctor Toscano Durán
vtoscano@us.es

Applied Mathematics I Department, University of Seville
Seville, Spain

July 18, 2024

Application of the representative measure approach to assess the reliability of decision trees in dealing with unseen vehicle collision data

Joint work with: Javier Perera Lago, Eduardo Paluzo Hidalgo, Sara Narteni and Matteo Rucco.

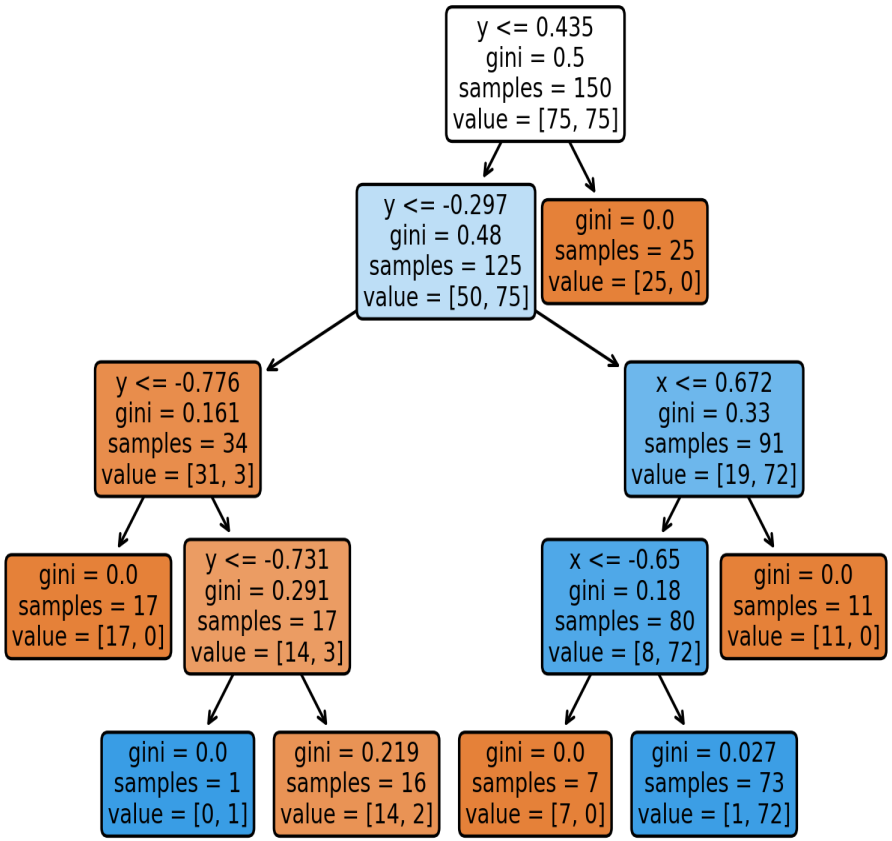
Take home message

- The importance of representative datasets in ML
- Assessing dataset similarity using ε-representativeness method for binary decision trees
- Significant correlation between ε-representativeness and feature importance ordering

Outline

Binary decision trees

Composition



- A **binary decision tree** (DT) is a rooted tree representing a partition of the feature space and is composed of **nodes** and **branches**.

XGBoost

- XGBoost is an ensemble of decision trees, i.e., it combines the predictions of several binary DTs that are sequentially built, each correcting errors of the previous one. Samples that are incorrectly predicted have a higher weight in the following trees.

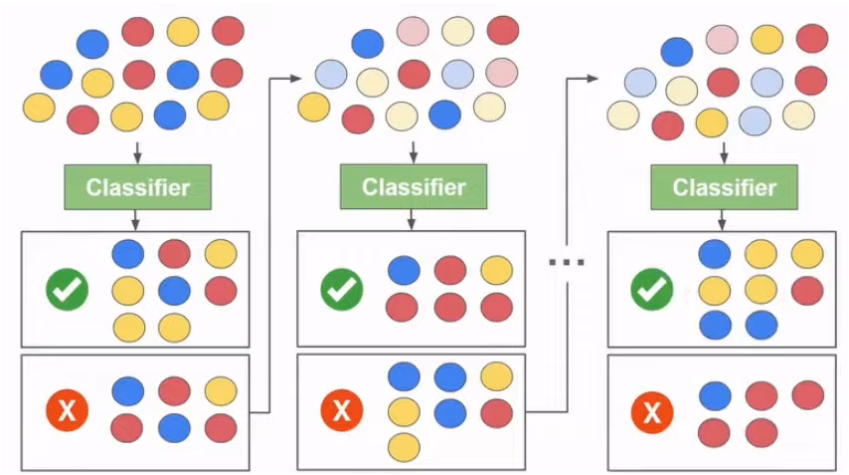


Figure: XGBoost process

Feature and Splitting condition

1. The feature and splitting condition are chosen to minimize node impurity.
2. Node impurity measures class label homogeneity: maximum when labels are even, minimum (pure nodes) when all labels are identical.
3. Information Gain measures how much the impurity is reduced due to the split.

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:

$$FI(j) = \sum_{\substack{i \in I \\ j_i = j}} N_i \cdot IG(n_i) \tag{1}$$

where N_i is the number of examples from data reaching the node n_i and $IG(n_i)$ is the information gain of node i .

Feature importance ordering

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

$$\mathrm{Sim}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left| \mathrm{pos}_x(f_i) - \mathrm{pos}_y(f_i) \right| \tag{2}$$

[1] "Barrera-Vicent, A., Paluzo-Hidalgo, E., Gutiérrez-Naranjo, M.A.: The metric-aware kernel-width choice for LIME"

Feature importance ordering

Suppose we have three subsets of features ordered based on their importance:

- x = [1,**2**,3,4]
- y = [1,**2**,4,3]
- z = [3,1,**2**,4]

Then $\text{Sim}(x, y) = 0.5$, $\text{Sim}(x, z) = 1$ and $\text{Sim}(y, z) = 1.5$.

The lower the value of this metric, the greater the similarity between the importance vectors.

ε-representativeness

Introducing ε-representativeness

Definition. 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.[2]

[2] "Gonzalez-Diaz, R., Gutiérrez-Naranjo, M.A., Paluzo-Hidalgo, E.: Topology-based representative datasets to reduce neural network training resources"

Illustrative example of ε-representativeness calculation

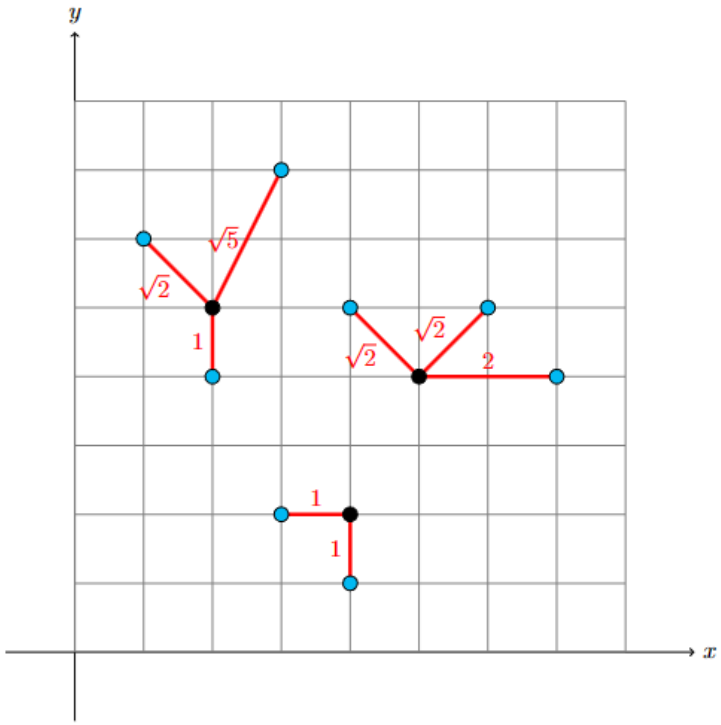


Figure: Calculation example of the ε-representativeness of a reduced dataset in relation to a larger one. It involves finding the closest representative point (black) for each original point (blue) and computing the distance. The maximum of these minimum distances, illustrated in red in the graphic,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

γ-balanced

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 $\tilde{x} \in \tilde{X}$.

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

$$\mathrm{Acc}(T, (X, \lambda_X)) = \mathrm{Acc}(T, (\tilde{X}, \lambda_{\tilde{X}})) \tag{3}$$

Experiments

Introduction

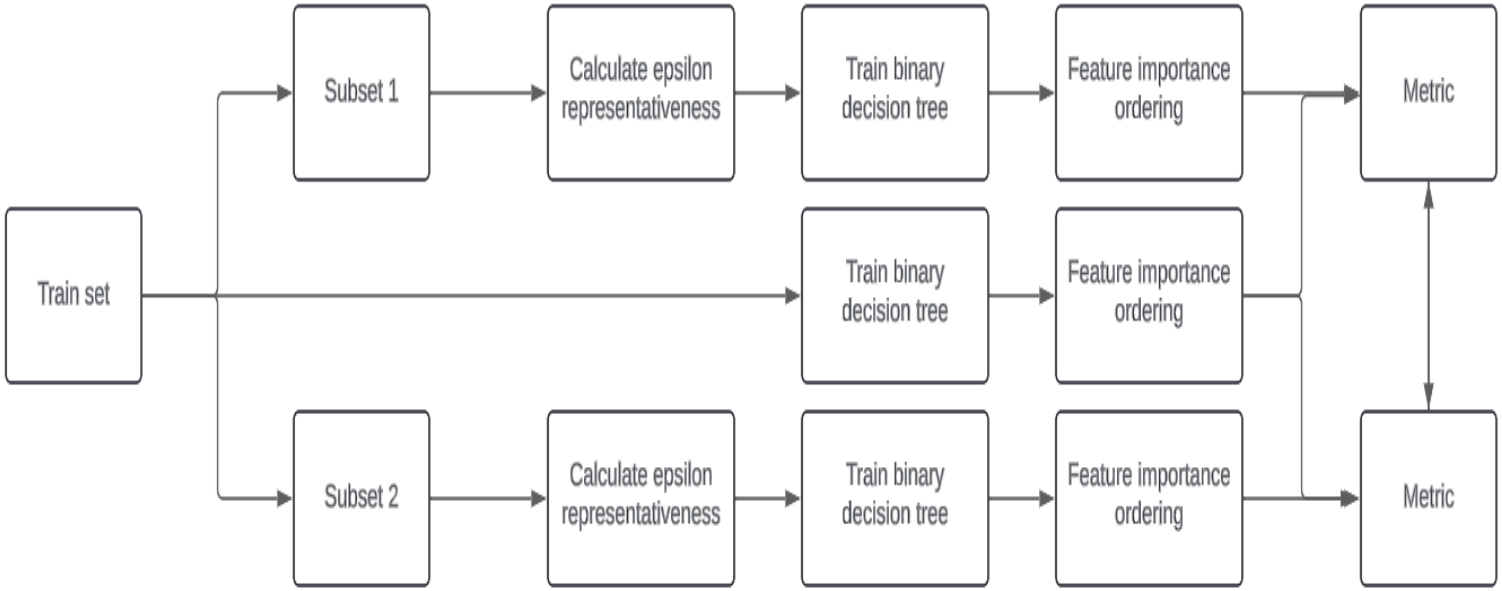


Figure: Methodology.

- The code for the experiments is available in a GitHub repository [1].
- Firstly, a 2D synthetic dataset was used as an illustration and then we extend to vehicle platooning dataset.

[1] https://github.com/Cimagroup/Application_Representative_Measure_Reliability_DT

2D Synthetic dataset

The dataset comprises 200 data distributed in two noisy concentric circles representing distinct classes as shown in next Figure.

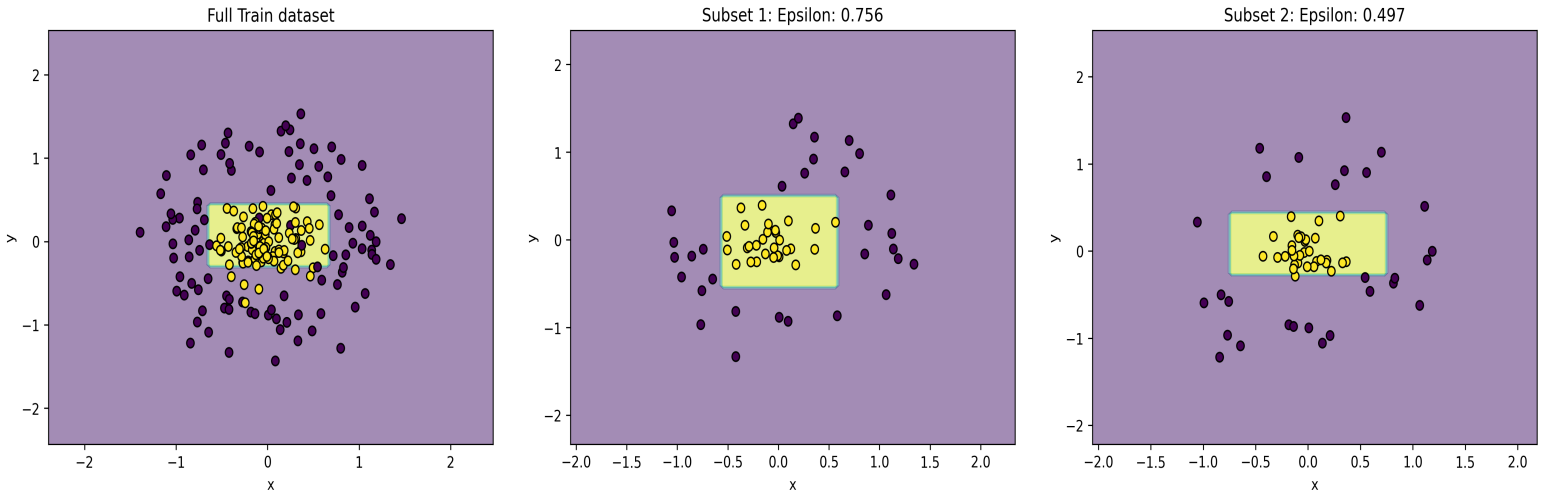


Figure: Full synthetic dataset generated using Scikit-learn and the two random subsets. From left to right: (1) the training set; (2) a subset composed of a 40% of the training set and $\varepsilon = 0.756$; (3) a subset composed of 40% of the training set and $\varepsilon = 0.497$. We can also see the decision boundaries of the binary DTs trained using each set of data.

2D Synthetic dataset

The accuracy values for the binary DTs on the test set were: 0.84 for the train dataset, 0.94 for Subset 1, and 0.82 for Subset 2.

Table: Feature importance percentage for the training set, and subset 2. We can see that the most important feature for the training set and Subset 2 is the same.

Data	x_1	x_2	ε
Training set	40.4	**59.6**	-
Subset 1	**50.37**	49.62	0.756
Subset 2	18.4	**81.6**	**0.497**

Vehicle Platooning dataset

This dataset consists of predicting whether a platoon of vehicles will collide based on features such as the number of cars or their speed. It is composed of 107,210 data with 23 numerical features.

Vehicle Platooning dataset

- ε-representativeness
  - Subset 1: 0.539
  - Subset 2: 0.655
- Better feature importance similarity with lower ε
  - Subset 1: 1.74
  - Subset 2: 1.83
- Significant correlation between ε and feature importance similarity. $Sp = 0.51$, p-value$= 5.2 \times 10^{-8}$

Vehicle Platooning dataset, extension to XGBoost

- ε-representativeness
  - Subset 1: 0.539
  - Subset 2: 0.655
- Better feature importance similarity with lower ε
  - Subset 1: 0.696
  - Subset 2: 1.823
- Significant correlation between ε and feature importance similarity. $Sp = 0.673$, p-value$= 1.79 \times 10^{-14}$

Conclusion

Conclusion

- Proved that similar accuracy is obtained under certain conditions of representativeness with binary decision trees (Theorem 1).
- Experiments shows significant correlation between feature importance ordering and ε-representativeness. According to our results, representative sets produce similar explanations of the dataset.

Future work

- Aim to provide theoretical guarantees regarding the ordering of feature importance.
- Plan distance-based comparison between the decision rules of binary decision trees.

Acknowledgement

- Maurizio Mongelli and Miguel A. Gutierrez-Naranjo for the insightful discussions and ideas.
- European Union HORIZON-CL4-2021-HUMAN-01-01 under grant agreement 101070028 (REXASI-PRO).
- TED2021-129438B-I00/ AEI/10.13039/501100011033 / Unión Europea NextGenerationEU/PRTR.

Thanks for your time! :)