Efficient Training of Robust Decision Trees Against Adversarial Examples

Vos, D.A.; Verwer, S.E.

Publication date
2021

Document Version
Final published version

Published in
BNAIC/BeneLearn 2021

Citation (APA)
Vos, D. A., & Verwer, S. E. (2021). Efficient Training of Robust Decision Trees Against Adversarial Examples. In E. L. A. Leiva, C. Pruski, R. Markovich, A. Najjar, & C. Schommer (Eds.), *BNAIC/BeneLearn 2021: 33rd Benelux Conference on Artificial Intelligence and 30th Belgian-Dutch Conference on Machine Learning* (pp. 702-703)

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Efficient Training of Robust Decision Trees Against Adversarial Examples

Daniël Vos and Sicco Verwer

Delft University of Technology, Delft, The Netherlands
{d.a.vos,s.e.verwer}@tudelft.nl

1 Introduction

Recently it has been shown that many machine learning models are vulnerable to adversarial examples: perturbed samples that trick the model into misclassifying them. Neural networks have received much attention but decision trees and their ensembles achieve state-of-the-art results on tabular data, motivating research on their robustness. Recently the first methods have been proposed to train decision trees and their ensembles robustly [4, 3, 2, 1] but the state-of-the-art methods are expensive to run.

We propose GROOT, an efficient algorithm for training robust decision trees. Like Chen et al. [3], we closely mimic the greedy recursive splitting strategy that traditional decision trees use and we score splits with the adversarial Gini impurity. We prove that the adversarial Gini impurity is concave with respect to the number of modifiable data points and use its analytical solution to compute the function in constant time. Our results show that GROOT trains trees 3 to 6 orders of magnitude faster than the state-of-the-art method TREANT [2] and trains random forests 100-1000 times faster than provably robust boosting [1].

2 GROOT: Growing Robust Trees

We introduce GROOT, an algorithm that trains decision trees that are robust against adversarial examples generated from a user-specified threat model. Like regular decision tree learning algorithms, GROOT runs in $\mathcal{O}(n \log n)$ time in terms of n samples. Similar to these algorithms, GROOT greedily makes splits according to a heuristic and while such strategies perform well in practice, they have no provable bound [5]. Where regular tree learning algorithms use the Gini impurity to score splits, GROOT uses the adversarial Gini impurity. This function represents the worst-case impurity after adversarial attacks, see Fig. 1.

3 Results

We evaluated the robustness of the algorithms on 13 tabular datasets by attacking all samples within a radius of 10% of the feature range in Fig. 2. GROOT decision trees and random forests on average perform as well as the existing

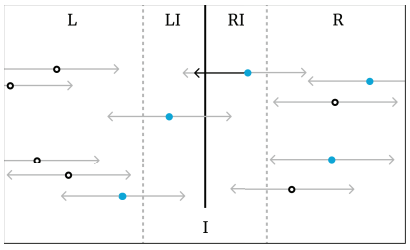

Fig. 1. We want to move samples from I over the threshold to maximize the weighted average of Gini impurities. Here we can move the single blue sample from RI into LI to maximize it.

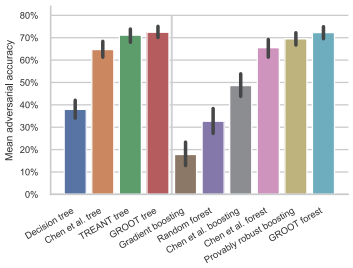

Fig. 2. Average adversarial accuracy over 13 structured datasets, GROOT trees and random forests achieve top results.

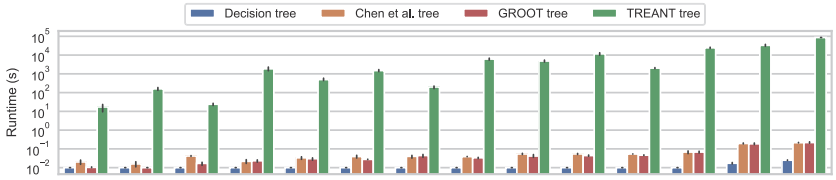

Fig. 3. Logarithmic training runtimes for single decision trees on different datasets. GROOT and Chen et al. run orders of magnitude faster than TREANT.

state-of-the-art works in trees and ensembles. To compare the efficiency of the algorithms, we plot the run times of each run in Figure 3 averaged over 5 data folds. All experiments ran on a single core of a Linux machine with 16 Intel Xeon CPU cores and 72GB of RAM total. Our results show that GROOT fits trees within seconds and scores as well as existing work against a box-shaped attack model. GROOT is available as a Scikit-learn compatible classifier[1].

References

1. Andriushchenko, M., Hein, M.: Provably robust boosted decision stumps and trees against adversarial attacks. arXiv preprint arXiv:1906.03526 (2019)
2. Calzavara, S., Lucchese, C., Tolomei, G., Abebe, S.A., Orlando, S.: Treant: Training evasion-aware decision trees. Data Mining and Knowledge Discovery pp. 1–31 (2020)
3. Chen, H., Zhang, H., Boning, D., Hsieh, C.J.: Robust decision trees against adversarial examples. In: ICML. pp. 1122–1131 (2019)
4. Kantchelian, A., Tygar, J.D., Joseph, A.: Evasion and hardening of tree ensemble classifiers. In: ICML. pp. 2387–2396 (2016)
5. Kearns, M.: Boosting theory towards practice: Recent developments in decision tree induction and the weak learning framework. In: Proceedings of the National Conference on Artificial Intelligence. pp. 1337–1339 (1996)

[1] https://github.com/tudelft-cda-lab/GROOT