Adaptive Anchor Data Augmentation for Robust Decision Trees

Hoin Jung [*1]

Abstract

Adversarial attacks pose a significant threat to machine learning models, especially those with complex decision boundaries such as decision trees. To enhance the robustness of decision trees against adversarial attacks, this paper proposes a novel data augmentation approach, *Adaptive Anchor Data Augmentation* (AADA). AADA uses anchors with adaptive radii as reference points to guide the decision tree to learn smoother decision boundaries. We evaluate the effectiveness of AADA through experiments on six benchmark datasets and five baseline models, showing that decision trees trained with our framework achieve better accuracy on adversarial samples while maintaining high performance on the original data. Furthermore, we demonstrate that our approach improves the smoothness of decision boundaries, as measured by the local Lipschitzness. Our results suggest that AADA is an effective strategy to enhance the robustness of decision trees against adversarial attacks while maintaining high accuracy on the original data.

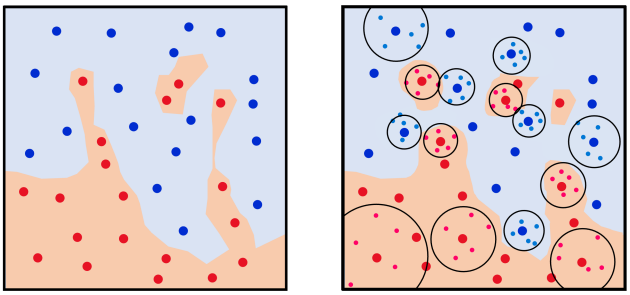

Figure 1. (a) Complex decision boundaries for decision trees. (b) Decision boundaries with the proposed method. The different color means different class label, and the colored background means the decision region separated by decision boundaries. In figure (b), arbitrary samples are chosen as center points of anchor. Each anchor consists of p additional data points distributed uniformly as ball-shape from center point. The radius of each ball is determined by the half of the minimum distance between a center point and the closest point from the center having different label.

1. Introduction

Decision trees have been widely used in various fields, such as finance, healthcare, and engineering, due to their interpretability and ease of use. One of the key advantages of decision trees is their sensitivity in catching complex decision boundaries. Unlike linear models, decision trees can capture nonlinear relationships between features, allowing them to handle complex datasets more effectively. Furthermore, decision trees still outperform deep learning models for tabular datasets, despite the recent advances in deep learning. However, a drawback of decision trees is their vulnerability to adversarial attacks. Since decision trees create complex decision boundaries, they are more susceptible to adversarial samples that are specifically designed to mislead the classifier. Therefore, many suggestions have been made to improve the robustness of decision trees, such as examples In this paper, we propose a data augmentation approach, *Adaptive Anchor Data Augmentation* (AADA), to enhance the robustness of decision trees against adversarial attacks. The anchors are created around randomly selected samples as center points, where the radius of each anchor is determined by interclass distance; the distance to the closest point of the opposite label. The anchors guide the decision tree to learn smoother decision boundaries, which are less likely to be affected by adversarial samples.

We evaluate the effectiveness of AADA through experiments on several benchmark datasets. The experimental results show that decision trees trained with our framework achieve better robustness against adversarial attacks, while maintaining high accuracy on the original data. Moreover, we compare the smoothness of decision boundaries between models with and without our approach.

In conclusion, our proposed approach provides a practical and effective way to enhance the robustness of decision trees against adversarial attacks. While there is no universal method to improve the robustness of decision trees, our approach offers a new perspective on data augmentation that leverages the power of anchors to create smoother decision boundaries.

2. Related Work

Decision trees have been widely used in various applications(Brutzkus et al., 2020; Blanc et al., 2020), including the medical field(Azar & El-Metwally, 2013; Lavanya & Rani, 2011), due to their interpretable nature and ability to handle both categorical and numerical features(Fiat & Pechyony, 2004). Recent works have proposed decision tree-based methods for tasks such as diagnosis prediction and treatment recommendation(Shehab et al., 2022). Moreover, decision trees have shown better performance than deep learning models on some tabular datasets(Shwartz-Ziv & Armon, 2022; Grinsztajn et al., 2022).

However, decision trees are vulnerable to adversarial attacks that manipulate the input data to mislead the model's prediction. Several works have proposed decision tree-specific attack methods, such as (Cheng et al., 2020; Papernot et al., 2016) and (Kantchelian et al., 2016), which can generate adversarial samples that cause misclassification. To mitigate this issue, many robust decision trees are suggested such as BBM-RS(Moshkovitz et al., 2021), GROOT(Vos & Verwer, 2021), GBDT(Chen et al., 2019a), and ROCT(Vos & Verwer, 2022) that are more resilient to adversarial attacks.

Data augmentation is a widely used technique to enhance the generalization ability of machine learning models. Several works have proposed data augmentation techniques for decision trees, such as(Chawla et al., 2002), (Tanaka & Aranha, 2019), and (Ionescu et al., 2022). These methods can improve the model's performanc in terms of accuracy. However, to the best of our knowledge, no existing work has focused on using data augmentation to enhance the robustness of decision trees against adversarial attacks.

3. Preliminaries

3.1. Zeroth Order Optimization Attack

Zeroth Order Optimization(ZOO) attack(Chen et al., 2017) is a black-box attack which can be adopted to any classifiers by using approximate gradient with a finite difference method. The untargeted objective function is defined as

$$f(\mathbf{x}) = \max\{\log[F(\mathbf{x})]_{t_0} - \max_{i \neq t_0} \log[F(\mathbf{x})]_i, -\kappa\} \quad (1)$$

where F is the output of a classifier, t_0 is the true class label for $\mathbf{x}$, and $\max_{i \neq t_0} \log[F(\mathbf{x})]_i$ represents the largest probability among other classes. Eq.1 can be optimized by any optimizer such as SGD, ADAM(Kingma & Ba, 2014), or Newton's method.

To obtain the gradient and Hessian estimate, Chen et al. (2017) adopts the symmetric difference quotient.

$$\hat{g}_i := \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i - f(\mathbf{x} - h\mathbf{e}_i)}{2h} \quad (2)$$

$$\hat{h}_i := \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_{ii}^2} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i - 2f(\mathbf{x}) + f(\mathbf{x} - h\mathbf{e}_i)}{h^2} \quad (3)$$

where h is a small constant, $h = 0.0001$ and e_i is a standard basis vector. Although ZOO attack is simple and slower than first-order method, it's sufficient to attack with very high success rate(Liu et al., 2020). Moreover, as the number of class label is two in binary classification, there's no big difference between the targeted and untargeted ZOO attack. Chen et al. (2017) suggested that ZOO attack with ADAM outperforms that with other optimizer.

4. Proposed Method

4.1. Adaptive Anchor

We create Adaptive Anchors as a data augmentation for train data $\mathbf{X} \in \mathbb{R}^{N \times d}$ to achieve robust decision trees against adversarial attack. n_c number of arbitrary data points $\mathbf{x}_{c,i}$ $\forall i = \{1, \cdots, n_c\}$ are selected as a center of an anchor, which consists of p uniformly distributed data points for each anchor with radius r_i, i.e. ℓ_2-ball such that

$$\mathcal{A}(\mathbf{x}_{c,i}, r_i) = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} \sim U(\|\mathbf{x} - \mathbf{x}_{c,i}\|_2 \leq r_i)\} \quad (4)$$

where $|\mathcal{A}(\mathbf{x}_{c,i}, r_i)| = m$. The data points in each anchor have the same label with the center. The radius r_i is adaptively determined by the half of the minimum interclass distance d_i, i.e. distance between a center of anchor and the closest point having a different class with the center point,

$$r_i = \frac{1}{2} d_i = \frac{1}{2} \min_{y_i \neq y_j} (\|\mathbf{x}_{c,i} - \mathbf{x}_j\|_2), \forall \mathbf{x}_j \in \mathbf{X} \quad (5)$$

In short, 5 allows the radius of anchor near the decision boundary be small, and the one far from the decision boundary be large, making a classifier be less sensitive to adversarial attack.

4.2. Training with Adaptive Anchor against adversarial attack

In training step, we choose $\gamma|\mathbf{X}|$ number of samples to make them the centers of anchor, i.e. $n_c = \gamma|\mathbf{X}|$ where γ is a hyperparameter determining the ratio of center point. Figure.1 shows the concept of the proposed AADA.

A decision tree $f(\mathbf{X})$ is trained with augmented data $\mathbf{X}'$ such that

$$\mathbf{X}' = \mathbf{X} \cup \bigcup_{i=1}^{n_c} \mathcal{A}(\mathbf{x}_{c,i}, r_i). \quad (6)$$

Based on the trained $f(\mathbf{X})$, an adversarial attack is applied to test dataset, $\mathbf{X}_{test}$ to create adversarial test samples to

Algorithm 1 Adaptive Anchor Data Augmentation against Adversarial Attack

Require: Training set $(\mathbf{X}, \mathbf{Y})$, hyperparameter γ and p, Adversarial attacker
Ensure: Prediction result on test dataset $\hat{\mathbf{Y}}_{test}$
$n_c \leftarrow \gamma|\mathbf{X}|$
for $\imath \in \{1, \cdots, n_c$ **do**
$r_i \leftarrow \frac{1}{2}\min_{y_i \neq y_j}(\|\mathbf{x}_{c,i} - \mathbf{x}_j\|_2), \forall \mathbf{x}_j \in \mathbf{X}$.
$\mathcal{A}(\mathbf{x}_{c,i}, r_i) \leftarrow \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} \sim U(\|\mathbf{x} - \mathbf{x}_{c,i}\|_2 \leq r_i)\}$, where $|\mathcal{A}(\mathbf{x}_{c,i}, r_i)| = p$.
end for
$\mathbf{X}' \leftarrow \mathbf{X} \cup \bigcup_{i=1}^{n_c} \mathcal{A}(\mathbf{x}_{c,i}, r_i)$.
Train a classifier f with augmented training dataset $(\mathbf{X}', \mathbf{Y}')$.
$\tilde{\mathbf{X}}_{test} \leftarrow \text{Attack}(\mathbf{X}_{test}, \mathbf{Y}_{test}, f(\mathbf{X}))$
$\hat{\mathbf{Y}}_{test} \leftarrow f(\tilde{\mathbf{X}}_{test})$

verify the robustness of the classifier such that

$$\tilde{\mathbf{X}}_{test} = \mathbf{X}_{test} + \delta, \tag{7}$$

where $\delta \in \mathbb{R}^{N_{test} \times d}$ is perturbation for $\mathbf{X}_{test}$ trained by adversary,

$$\delta = \text{Attack}(\mathbf{X}_{test}, \mathbf{Y}_{test}, f(\mathbf{X})). \tag{8}$$

The overall algorithm are shown in Algorithm.1.

4.3. Local Lipschitzness

To evaluate the robustness of a classifier, we adopt L-Local Lipschitzness as an evaluation metric(Hein & Andriushchenko, 2017; Yang et al., 2020).

Definition 4.1. A function $g : \mathcal{X} \to \mathbb{R}$ is L-locally Lipschitz around a center point $\mathbf{x} \in \mathcal{X}$ with radius r if $d(\mathbf{x}, \mathbf{x}') \leq r$ and $d\big(g(\mathbf{x}), g(\mathbf{x}')\big) \leq L \cdot d(\mathbf{x}, \mathbf{x}')\ \forall \mathbf{x}$ for a constant $L \geq 0$.

Hein & Andriushchenko (2017) proved that local Lipschitzness guarantees robustness which means the classifier's prediction does not change in a certain range of ball from the center point for any types of adversarial samples, transformation, or noise. Yang et al. (2020) suggested an evaluation metric to measure the robustness of classifier using *average Local Lipschitzness* such that

$$\hat{Lip}_\epsilon = \frac{1}{n}\sum_{i=1}^{n} \max_{\mathbf{x}_i' \in \mathbb{B}_{inf}(\mathbf{x}_i, \epsilon)} \frac{\|f(\mathbf{x}_i) - f(\mathbf{x}_i')\|_1}{\|\mathbf{x}_i - \mathbf{x}_i'\|_\infty}. \tag{9}$$

where ϵ is user-specific hyperparameter. Yang et al. (2020) estimate $\hat{Lip}$ iteratively with projected gradient descent using step size $\epsilon/5$.
In this paper, we define the final evaluation metric

$$\hat{Lip} = \frac{1}{|\mathcal{E}|}\sum_{\epsilon \in \mathcal{E}} \hat{Lip}_\epsilon \tag{10}$$

where $\epsilon \in \mathcal{E} = \{0.01, 0.02, \cdots, , 0.99, 1.0\}$.

5. Experimental Result

5.1. Dataset

To evaluate the effectiveness of our proposed approach, we select six publicly available datasets from different domains. These datasets are commonly used in machine learning research and have been previously used to evaluate the robustness of models against adversarial attacks(Andriushchenko & Hein, 2019; Moshkovitz et al., 2021; Vos & Verwer, 2021). Especially, we choose healthcare related dataset only to show that AADA can be used for constructing robust decision trees against adversarial attack on patient data.

Drug Consumption Dataset. Drug Consumption dataset (Dua et al., 2017) contains records from 1,885 respondents about drug consumption. Each data point has 12 attributes including the level of education, age, gender, and so on. The original task is multi classification for 7 classes of whether and when respondents experienced drugs, but our prediction goal is abridged whether they consumed cocaine or not.

Heart Disease Dataset. Heart Disease dataset contains 303 instances with 14 attributes such as age, sex, and numerical values about heart disease. The goal is to predict a patient has a heart disease.

Others. Diabetes, Mammography, Breast Cancer, and Ionosphere datasets are provided by OpenML repository. Pima Indians Diabetes dataset consists of 768 instances with 8 attributes, aiming to predict whether a patient has a diabetes or not. Wisconsin Breast Cancer dataset consists of 699 instances with 11 context cytology features. It can be used to predict breast cancer from cytology features. Mammography dataset is also used to predict breast cancer with 11,183 instances with 7 features. Ionosphere dataset is not a medical related data, but still helpful to be used as a benchmark tabular data, consisting of 351 instances with 35 features.

Drug		CART	AdaBoost	GradientBoosting	RandomForest	XGBoost	GROOT
Baseline	Acc.	0.6090 ± 0.0339	0.6807 ± 0.0475	0.6897 ± 0.0436	0.6891 ± 0.0333	0.6775 ± 0.0407	0.6600 ± 0.0412
	Acc. on adv.	0.4812 ± 0.0279	0.3575 ± 0.0426	0.3289 ± 0.0389	0.5141 ± 0.0392	0.5088 ± 0.0301	0.5831 ± 0.0409
	Local Lip.	11.9611 ± 1.4933	8.0575 ± 0.7961	8.2251 ± 0.8396	**8.2968 ± 0.6042**	10.4016 ± 0.5916	53.3614 ± 4.7138
Anchor	Acc.	0.6249 ± 0.0363	0.6790 ± 0.0391	0.6838 ± 0.0427	0.6849 ± 0.0343	0.6764 ± 0.0435	-
	Acc. on adv.	**0.6127 ± 0.0365**	**0.6000 ± 0.0330**	**0.5698 ± 0.0262**	**0.5575 ± 0.0287**	**0.5592 ± 0.0255**	-
	Local Lip.	**11.7922 ± 1.2003**	**7.5474 ± 1.3817**	**7.3198 ± 0.9736**	8.3308 ± 0.6353	**10.2115 ± 0.6635**	-

Heart		CART	AdaBoost	GradientBoosting	RandomForest	XGBoost	GROOT
Baseline	Acc.	0.7111 ± 0.0934	0.7963 ± 0.0727	0.8148 ± 0.0741	0.8111 ± 0.0651	0.7926 ± 0.0744	0.7481 ± 0.0857
	Acc. on adv.	0.6444 ± 0.0646	0.4852 ± 0.1079	0.4926 ± 0.0664	**0.7185 ± 0.0707**	**0.6630 ± 0.0452**	0.6741 ± 0.0919
	Local Lip.	**0.0222 ± 0.0294**	0.0954 ± 0.1588	0.0549 ± 0.0584	0.0531 ± 0.0676	**0.033 ± 0.0475**	0.3238 ± 0.1065
Anchor	Acc.	0.7222 ± 0.0880	0.8111 ± 0.0749	0.7926 ± 0.0667	0.7778 ± 0.0597	0.7889 ± 0.0760	-
	Acc. on adv.	**0.7185 ± 0.1010**	**0.7296 ± 0.0923**	**0.7000 ± 0.0767**	0.6407 ± 0.0845	0.6259 ± 0.0607	-
	Local Lip.	0.035 ± 0.0491	**0.0318 ± 0.0622**	**0.0393 ± 0.0617**	**0.0190 ± 0.0306**	**0.033 ± 0.0475**	-

Diabates		CART	AdaBoost	GradientBoosting	RandomForest	XGBoost	GROOT
Baseline	Acc.	0.6798 ± 0.0563	0.7475 ± 0.0508	0.7696 ± 0.0457	0.7632 ± 0.0545	0.7201 ± 0.0548	0.7228 ± 0.0568
	Acc. on adv.	0.5195 ± 0.0280	0.2563 ± 0.0513	0.2369 ± 0.0445	0.3683 ± 0.0518	0.4141 ± 0.0558	0.4780 ± 0.0693
	Local Lip.	56.4847 ± 11.8601	**40.7800 ± 9.6008**	**38.8410 ± 7.9452**	43.7244 ± 9.7973	49.6119 ± 9.1651	122.1606 ± 19.8199
Anchor	Acc.	0.6991 ± 0.0414	0.7489 ± 0.0483	0.7683 ± 0.0478	0.7657 ± 0.0456	0.7383 ± 0.0597	-
	Acc. on adv.	**0.6405 ± 0.0673**	**0.5419 ± 0.0778**	**0.5052 ± 0.0498**	**0.4649 ± 0.0513**	**0.5104 ± 0.0438**	-
	Local Lip.	**53.2102 ± 9.7323**	43.9431 ± 13.2394	41.2516 ± 10.0705	**43.0882 ± 11.0314**	**47.3415 ± 10.0376**	-

Mammo		CART	AdaBoost	GradientBoosting	RandomForest	XGBoost	GROOT
Baseline	Acc.	0.7856 ± 0.0362	0.8002 ± 0.0413	0.8033 ± 0.0380	0.7877 ± 0.0433	0.7867 ± 0.0372	0.8044 ± 0.0378
	Acc. on adv.	0.7003 ± 0.0465	**0.7732 ± 0.0408**	**0.7327 ± 0.0495**	**0.7576 ± 0.0345**	**0.7815 ± 0.0305**	**0.7940 ± 0.0435**
	Local Lip.	**7.1066 ± 12.4130**	**32.2620 ± 52.4640**	**4.6179 ± 10.0536**	**7.8107 ± 12.3759**	**5.6878 ± 10.3242**	361.3532 ± 28.1073
Anchor	Acc.	0.7814 ± 0.0342	0.7908 ± 0.0338	0.7877 ± 0.0308	0.7793 ± 0.0359	0.7773 ± 0.0346	-
	Acc. on adv.	**0.7440 ± 0.0542**	0.6868 ± 0.0595	0.7066 ± 0.0572	0.6160 ± 0.0428	0.7005 ± 0.0687	-
	Local Lip.	17.7274 ± 38.0165	36.1674 ± 52.8626	15.947 ± 38.3866	17.7274 ± 38.0165	17.7274 ± 38.0165	-

Cancer		CART	AdaBoost	GradientBoosting	RandomForest	XGBoost	GROOT
Baseline	Acc.	0.9342 ± 0.0278	0.9605 ± 0.0171	0.9708 ± 0.0160	0.9707 ± 0.0197	0.9678 ± 0.0224	0.9473 ± 0.0209
	Acc. on adv.	0.8755 ± 0.0336	0.1759 ± 0.0516	0.2371 ± 0.0781	0.5052 ± 0.0927	**0.7380 ± 0.0470**	0.5334 ± 0.1030
	Local Lip.	80.2966 ± 119.8837	38.2902 ± 53.3038	**10.0234 ± 16.1989**	18.878 ± 30.0236	**19.7731 ± 23.6790**	424.4295 ± 223.8043
Anchor	Acc.	0.9429 ± 0.0221	0.9591 ± 0.0214	0.9591 ± 0.0274	0.9649 ± 0.0238	0.9708 ± 0.0196	-
	Acc. on adv.	**0.9370 ± 0.0227**	**0.6623 ± 0.1271**	**0.6517 ± 0.0802**	**0.7292 ± 0.0600**	0.6339 ± 0.1262	-
	Local Lip.	**33.0955 ± 31.4454**	**21.6268 ± 35.2081**	12.022 ± 17.7957	**9.5153 ± 12.7421**	25.5842 ± 26.8008	-

ionosphere		CART	AdaBoost	GradientBoosting	RandomForest	XGBoost	GROOT
Baseline	Acc.	0.8861 ± 0.0662	0.9144 ± 0.0528	0.9401 ± 0.0393	0.9287 ± 0.0692	0.9287 ± 0.0516	0.8689 ± 0.0575
	Acc. on adv.	0.7667 ± 0.1073	0.4846 ± 0.0800	0.4813 ± 0.065	0.5725 ± 0.0971	0.5164 ± 0.1127	0.5957 ± 0.0808
	Local Lip.	2.3735 ± 1.7265	1.0231 ± 0.6246	0.8008 ± 0.572	0.9712 ± 0.6302	0.9397 ± 0.6926	14.2667 ± 4.5570
Anchor	Acc.	0.9117 ± 0.0322	0.9174 ± 0.0563	0.9287 ± 0.0516	0.9344 ± 0.0528	0.9315 ± 0.0546	-
	Acc. on adv.	**0.8860 ± 0.0542**	**0.8263 ± 0.0756**	**0.6860 ± 0.0936**	**0.6890 ± 0.0655**	**0.6898 ± 0.0721**	-
	Local Lip.	**1.2335 ± 1.0496**	**0.9215 ± 0.5443**	**0.7023 ± 0.5855**	**0.8108 ± 0.5906**	**0.8100 ± 0.5902**	-

Table 1. Experimental Results for AADA with five baseline model and GROOT. The **bold** results mean the better adversarial accuracy and smoothness between with and without AADA for each baseline. The blue results mean the best result across the all reported outcome for each dataset.

5.2. Experimental Setting

We conduct extensive experiments to verify the enhancement of robustness with our proposed method. Five different models are used including Decision Tree(CART)(Breiman et al., 1984; Loh, 2011), AdaBoost(Freund & Schapire, 1995), GradientBoosting(Friedman, 2001), Random Forest(Ho, 1995), and XGBoost(Chen & Guestrin, 2016). We run 10 experiments for each case and report the mean and standard deviation. We set the center point ratio $\gamma = 0.1$ and the number of points in an anchor $p = 5$. We compare test accuracy and Local Lipschitzness on pure test samples and adversarial test samples for each classifier trained by pure training samples and augmented training samples respectively. The overall process of experiments is described in Figure.2

5.3. Result Analysis

We evaluated the effectiveness of our proposed Adaptive Anchor Data Augmentation (AADA) method on several decision tree models, including CART, AdaBoost, GradientBoosting, RandomForest, and XGBoost. The experimental results show that applying AADA to any baseline models improve their robustness against adversarial attacks generated by the ZOO attack. The adversarial accuracy of the models is significantly improved compared to their performance without AADA as shown in Figure.4 and Table.4.2.

In Figure.3, we visualize the decision boundaries and adversraial samples on Diabetes dataset. The visualization is

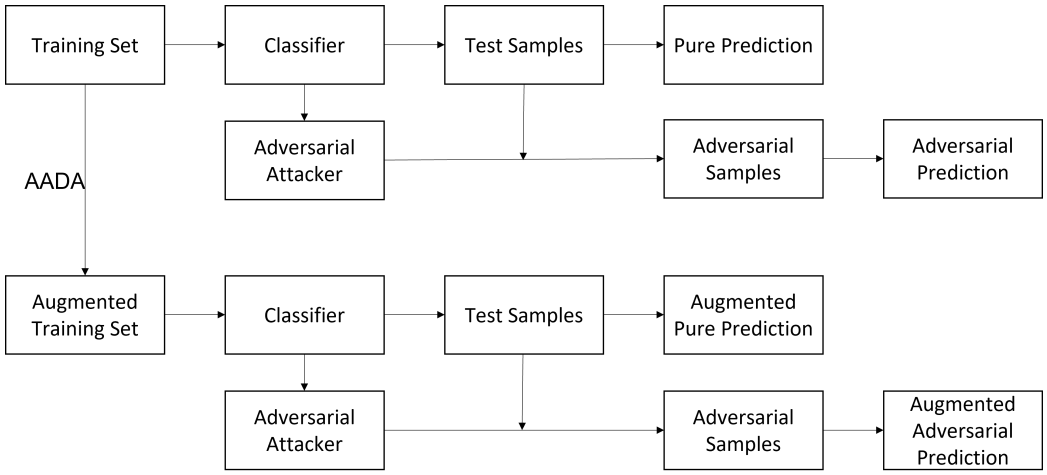

Figure 2. A classifier learns a training dataset and is tested on a test subset. An adversarial attack produces adversarial samples on the learned classifier perturbing the test samples. The test accuracy results on pure test samples and adversarial test samples are expected to be different if the adversarial attack succeeds. However, as we use AADA on the training dataset, the test accuracy gap between pure test samples and adversarial test samples will be decreased. Therefore, four different test accuracy will be compared to verify the ability to make the decision tree robust.

conducted only on two important features in dataset. The Figure.3 shows that the adversarial attack perturbed the test samples easily since the decision boundaries are too complex. On the other hand, the proposed method make the decision boundary smooth and make the adversarial attack harder. The importance score is extracted by XGBoost.

However, we also observe that the improvement in adversarial accuracy is not consistent across all datasets. In some cases, AADA does not result in significant improvements, which suggests that the effectiveness of the method may depend on the specific characteristics of the dataset and the baseline model. Nonetheless, our results demonstrate that AADA can be a promising approach to enhance the robustness of decision trees against adversarial attacks.

We also evaluate the smoothness of the decision boundaries of the models using Local Lipschitzness, which measures the maximum variation in the model's output with respect to small variations in the input. However, we found that improvements in Local Lipschitzness did not always correspond to improvements in adversarial accuracy. This suggests that Local Lipschitzness may not be a reliable indicator of a model's robustness against adversarial attacks, and other metrics may need to be considered such as (Chen et al., 2019b)

Finally, we compare the performance of AADA with another robust decision tree, GROOT(Vos & Verwer, 2021). Our results showed that AADA generally outperformed GROOT in terms of adversarial accuracy and smoothness of the decision boundaries, demonstrating the effectiveness of our proposed method.

5.4. Ablation Study

We compare the test accuracy reduction results varying the number of points $p = 5, 10, 20$ in each anchor and the center points $\gamma = 0.1, 0.2, \cdots, 0.5$ on GradientBoosting and AdaBoost as shown in Figure.5. The test accuracy reduction indicates the difference between test accuracy on pure samples and adversarial samples, where the lower reduction rate is better. The results shows that the number of points p in each anchor doesn't affect significantly. However, when too many center points are selected such as $\gamma \geq 0.4$, the reduction rate increase, which means the AADA doesn't work well.

6. Limitation

Currently, the AADA is limited to tabular dataset with tree-based model. AADA can be adopted to any combinations of data and models such as image on CNN and tabular dataset on MLP. However, if the decision boundary is not complex, the AADA might not be works well on Neural Networks. Moreover, the Local Lipschitzness doesn't look proper to reflect the robustness of a model. Fairer robustness metrics are required such as (Chen et al., 2019b). In addition to, only single attack method and robust model are considered at this moment. The extensive experiments on various attack method such as Cheng's attack, Papernot's attack and Kantchelian's attack and robust decision trees such as ROCT, BBM-RS are required. Lastly, more mathematical evidence is required to justify the ability of AADA enhancing robustness of decision tree.

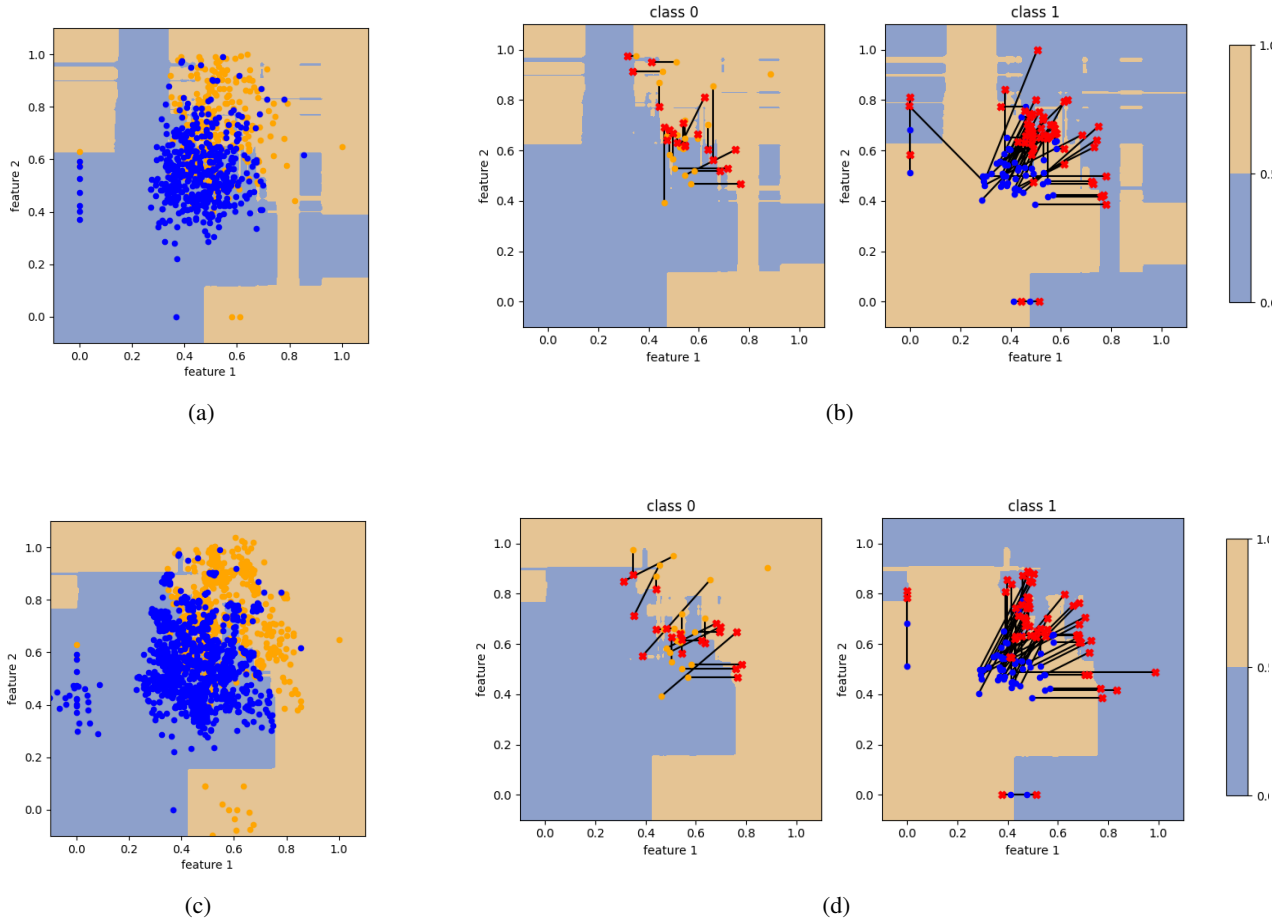

Figure 3. (a) Raw training samples and decision boundary (c) Augmented training samples and decision boundary (b),(d) test samples and adversarial samples. The upper figures show the decision boundary and adversarial samples on naive training samples, and the lower figures show the decision boundary for training samples with AADA. The different classes are indicated different colors, yellow and blue. The red cross points are the adversarial samples for corresponding original test samples. The left figures in (b) and (d) shows the adversarial attack for class 0 test points, and the right ones for class 1 test points. The decision boundaries in lower figures(with AADA) is smoother than the upper figures(naive).

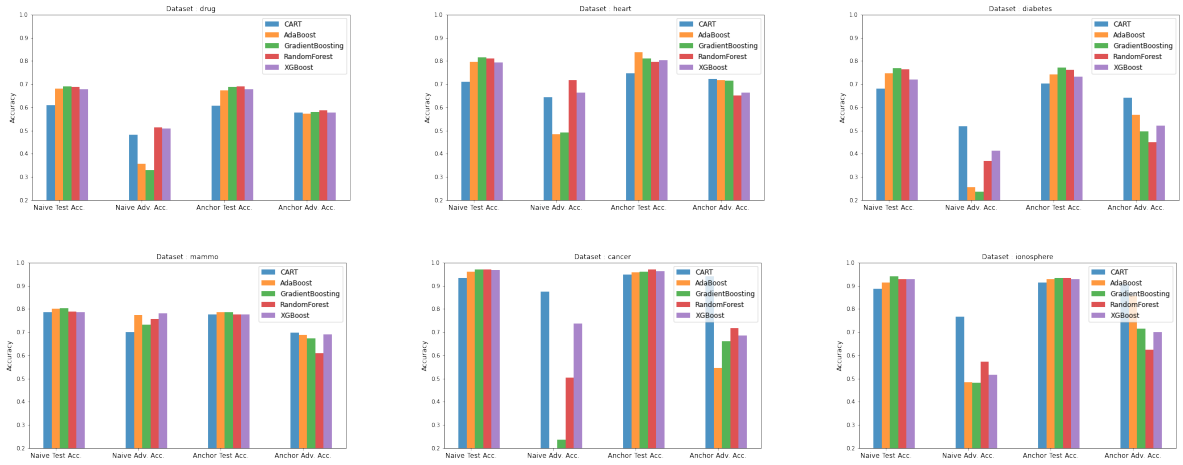

Figure 4. Test accuracy on pure test samples and adversarial test samples w/ and w/o AADA for each tree-based model. In most case, the accuracy reduction by adversarial attack decrease when the classifier is trained by augmented data.

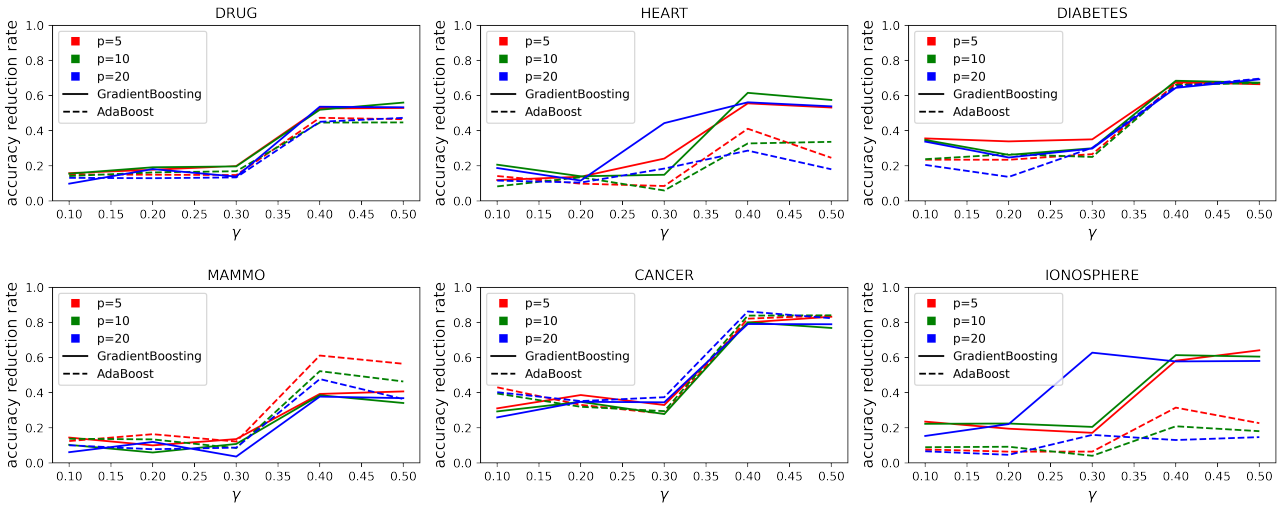

Figure 5. Ablation study results about test accuracy reduction which indicates the difference between test accuracy on pure samples and adversarial samples(lower is better) on six different dataset for two different models. The solid line means the results for GradientBoosting and the dashed line means the results for AdaBoost. Different color means the different number of points p in an anchor.

7. Conclusion

In this paper, we proposed a novel data augmentation approach, Adaptive Anchor Data Augmentation (AADA), to enhance the robustness of decision trees against adversarial attacks. Through experiments on six benchmark datasets and five baseline models, we showed that our approach significantly improves the accuracy on adversarial samples while maintaining high performance on the original data. Furthermore, we demonstrated that decision trees trained with our framework have smoother decision boundaries, as measured by the local Lipschitzness, compared to models without AADA.

Our results suggest that the use of anchors with adaptive radii is an effective strategy to guide the decision tree to learn smoother decision boundaries, which can improve the model's robustness against adversarial attacks. We believe that our approach can be applied to other types of models beyond decision trees, and to various real-world applications where the robustness of machine learning models is critical.

In summary, our proposed method, Adaptive Anchor Data Augmentation, provides a practical and effective way to enhance the robustness of decision trees against adversarial attacks, while maintaining high accuracy on the original data.

References

Andriushchenko, M. and Hein, M. Provably robust boosted decision stumps and trees against adversarial attacks. *Advances in Neural Information Processing Systems*, 32, 2019.

Azar, A. T. and El-Metwally, S. M. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23:2387–2403, 2013.

Blanc, G., Lange, J., and Tan, L.-Y. Provable guarantees for decision tree induction: the agnostic setting. In *International Conference on Machine Learning*, pp. 941–949. PMLR, 2020.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. Cart. *Classification and regression trees*, 1984.

Brutzkus, A., Daniely, A., and Malach, E. Id3 learns juntas for smoothed product distributions. In *Conference on Learning Theory*, pp. 902–915. PMLR, 2020.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Chen, H., Zhang, H., Boning, D., and Hsieh, C.-J. Robust decision trees against adversarial examples. In *International Conference on Machine Learning*, pp. 1122–1131. PMLR, 2019a.

Chen, H., Zhang, H., Si, S., Li, Y., Boning, D., and Hsieh, C.-J. Robustness verification of tree-based models. *Advances in Neural Information Processing Systems*, 32, 2019b.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Cheng, M., Yi, J., Chen, P.-Y., Zhang, H., and Hsieh, C.-J. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3601–3608, 2020.

Dua, D., Graff, C., et al. Uci machine learning repository. 2017.

Fiat, A. and Pechyony, D. Decision trees: More theoretical justification for practical algorithms. In *Algorithmic Learning Theory: 15th International Conference, ALT 2004, Padova, Italy, October 2-5, 2004. Proceedings 15*, pp. 156–170. Springer, 2004.

Freund, Y. and Schapire, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory: Second European Conference, EuroCOLT'95 Barcelona, Spain, March 13–15, 1995 Proceedings 2*, pp. 23–37. Springer, 1995.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*, 2022.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in neural information processing systems*, 30, 2017.

Ho, T. K. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pp. 278–282. IEEE, 1995.

Ionescu, A., Hai, R., Fragkoulis, M., and Katsifodimos, A. Join path-based data augmentation for decision trees. In *2022 IEEE 38th International Conference on Data Engineering Workshops (ICDEW)*, pp. 84–88. IEEE, 2022.

Kantchelian, A., Tygar, J. D., and Joseph, A. Evasion and hardening of tree ensemble classifiers. In *International conference on machine learning*, pp. 2387–2396. PMLR, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lavanya, D. and Rani, K. U. Performance evaluation of decision tree classifiers on medical datasets. *International Journal of Computer Applications*, 26(4):1–4, 2011.

Liu, S., Chen, P.-Y., Kailkhura, B., Zhang, G., Hero III, A. O., and Varshney, P. K. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

Loh, W.-Y. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

Moshkovitz, M., Yang, Y.-Y., and Chaudhuri, K. Connecting interpretability and robustness in decision trees through separation. In *International Conference on Machine Learning*, pp. 7839–7849. PMLR, 2021.

Papernot, N., McDaniel, P., and Goodfellow, I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., and Gandomi, A. H. Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458, 2022.

Shwartz-Ziv, R. and Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Tanaka, F. H. K. d. S. and Aranha, C. Data augmentation using gans. *arXiv preprint arXiv:1904.09135*, 2019.

Vos, D. and Verwer, S. Efficient training of robust decision trees against adversarial examples. In *International Conference on Machine Learning*, pp. 10586–10595. PMLR, 2021.

Vos, D. and Verwer, S. Robust optimal classification trees against adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8520–8528, 2022.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.