On the Robust Splitting Criterion of Random Forest

Bin-Bin Yang, Wei Gao, Ming Li
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
Email: {yangbb, gaow, lim}@lamda.nju.edu.cn

***Abstract*—Splitting criteria have played an important role in the construction of decision trees, and various trees have been developed based on different criteria. This work presents a unified framework on various splitting criteria from the perspective of loss functions, and most classical splitting criteria can be viewed essentially as the optimizations of loss functions in this framework. We further introduce a new splitting criterion, named *pairwise gain*, which is motivated from a lower bound on the mutual coupling of pairwise loss. Theoretically, we prove that this new criterion is robust to symmetric and asymmetric label noises simultaneously. Based on this new criterion, we develop another variant of random forests, and extensive experiments are provided to verify its robustness.**

***Index Terms*—decision tree, splitting criterion, random forest, label noise, robustness**

I. Introduction

Decision trees have been a standard algorithm in machine learning, computer vision, information retrieval, etc. From the pioneer works of CART [1] and C4.5 [2], various decision trees have been developed [3]–[6] during the past decades, and the advantages of decision trees include good predictive performance, computational efficiency, interpretability, etc. Recent years have also witnessed the increasing popularity on decision trees as the base learners for ensemble algorithms [4], [6], [7]. For example, Microsoft Kinect makes real-time human pose estimation from single depth images by trees trained on millions of examples [8].

The construction of decision trees can be viewed as a top-down greedy procedure [9]. Given a data set, we begin with a root node, and each split partitions the training set into left and right subsets by some test on each instance, and the highest-scoring split is selected based on some certain criterion. We then partition the data set accordingly and grow the tree with the two newly created child nodes. This procedure is applied recursively until some stopping conditions are reached, such as a maximum tree depth or minimum sample size [1]. In such a procedure, splitting criteria play an important role in generalization error and structure of the learned tree.

Different decision trees have been developed based on various splitting criteria. For example, Breiman et al. [1] introduced the classical tree CART based on gini impurity, which measures how often a random example would be misclassified according to the class distribution. Quinlan [10] proposed information gain as another criterion for decision trees from an information-theoretic view, which considers the mutual information between local node decision (left or right nodes) and predictive output. More splitting criteria have been introduced along this line, such as sum minority [11], DKM [12], etc., whereas there is a lack of understanding on different splitting criteria from the perspective of loss functions.

This work introduces a unified framework for previous splitting criteria and proposes a new criterion. The main contributions can be summarized as follows:

- We exploit the relationship between splitting criteria and loss functions in the unified framework, and find that many classical splitting criteria are essentially equivalent to the optimizations of loss functions.
- A new splitting criterion is proposed in our framework, called *pairwise gain*, motivated from the optimization of pairwise loss. We prove that the *pairwise gain* criterion is robust to symmetric and symmetric label noises.
- We develop new decision trees and random forests based on the proposed *pairwise gain*. Extensive experiments show its robustness in comparison with classical criteria, also with more compact trees and less running-time cost.

II. Related Work

Splitting criteria have played an important role in the structure and generalization error of the learned decision trees, as shown empirically in [13]–[15], and there also have been some theoretical studies that analyzed the properties of splitting criteria [16]–[19]. However, they focused on impurity measures like Shannon entropy and gini index to unify the view on splitting criteria since splitting criteria were regarded as weighted sums of two impurity measures. In this paper, we analyze splitting criteria from the perspective of loss functions.

In the work [7] and [20], the authors derived splitting criteria from the second-order approximation of the additive training loss for gradient tree boosting, whereas their work cannot derive the classical splitting criteria. In contrast, our unified framework includes previous common splitting criteria and is suitable for a single tree and tree ensemble.

Not many studies are known about the robustness of decision tree learning under label noises. It was observed that label noises in the training data increase the size of the learned tree and that detecting and removing noisy examples improves the learned tree [21]. Ghosh et al. [22] presented some theoretical analysis to show that many popular decision tree algorithms are robust to symmetric label noises under large sample size. In this paper, we propose a new splitting criterion whose robustness is proved theoretically under both symmetric and asymmetric label noises. We also empirically study our criterion for random forests, since bagging and random split selection are immune to label noises [4].

The rest of the paper is organized as follows: Section III introduces some preliminaries. In Section IV we present the unified framework, where previous splitting criteria essentially optimize different (pointwise) loss functions. In Section V we propose a new splitting criterion from pairwise loss and prove its robustness to label noises. Section V shows the empirical studies for our proposed criterion. We finally conclude with future work in Section VII.

III. Preliminaries

Let $\mathcal{X} \subset \mathbb{R}^d$ be the instance/input space. Let $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{Y} = \{1, 2, ..., K\}$ denote the output space for regression and classification, respectively. Suppose that $\mathcal{D}$ is an underlying distribution over the product space $\mathcal{X} \times \mathcal{Y}$. We can observe training data $S_m = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_m, y_m)\}$, where each example is drawn i.i.d. from the distribution $\mathcal{D}$.

Given a decision tree, an instance $\mathbf{x} \in \mathcal{X}$ is directed from the tree root to a leaf node via internal nodes. Each internal node performs a binary test by evaluating a split function $s(\mathbf{x})\colon \mathbb{R}^d \to \{0, 1\}$, i.e., the instance $\mathbf{x}$ is directed to left child node for $s(\mathbf{x}) = 0$; and right child otherwise. Finally, one leaf node presents the output for the instance $\mathbf{x}$.

Let S be the set of training examples in one given node. Without loss of generality, we assume

$$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\} \quad \text{for} \quad n \leq m.$$

Later in the paper, $\mathbb{I}[\cdot]$ denotes the indicator function, which returns 1 if the argument is true and 0 otherwise.

IV. The Framework based on loss functions

Most supervised learning tasks can be formulated to learn a function f by optimization of objective function as follows:

$$R(f) = \sum_{i=1}^{m} l(f(x_i), y_i),$$

where $l(\cdot)$ is a loss function such as square loss, etc.

Generally, it is difficult to directly optimize the objective function $R(f)$ for decision trees, due to the discrete and sequential nature of the decisions in a tree. In practice, a feasible solution is to use the greedy algorithm from the root, and then add the branches iteratively, as most algorithms of decision tree induction adopted.

Given some node, the objective function after splitting on this node can be written as

$$L(s, f_L, f_R) = \sum_{i:s(x_i)=0} l(f_L(x_i), y_i) + \sum_{i:s(x_i)=1} l(f_R(x_i), y_i),$$

where $s(\cdot)$ denotes a binary split function which decides whether an input instance that reaches this node should progress through the left or right branch emanating from the node; and $f_L(\cdot), f_R(\cdot)$ denote the output function w.r.t the left and right child, respectively.

Table I
The relationship between loss functions and splitting criterions

Task	Splitting criterion	Loss function
Classification	Gini Impurity	Square Loss
	Information Gain	Softmax Loss
	Sum Minority	0-1 Loss
	DKM	Exponential Loss
Regression	Variance Reduction	Square loss
	Mean Absolute Error	Absolute loss

Given a specific split function $s(\cdot)$, we have the reduction of objective function

$$t(s) = \min_{w} \sum_{i=1}^{n} l(w, y_i) - \min_{w^L} \sum_{i:s(x_i)=0} l(w^L, y_i) - \min_{w^R} \sum_{i:s(x_i)=1} l(w^R, y_i), \quad (1)$$

where w, w^L, w^R denote the output values of the node and left and right child node, since instances in one node have the same output values.

Based on this equation, we can directly prove that the essence of various splitting criteria is to optimize some loss functions greedily. For instance, the frequently-used criteria gini impurity and information gain are essentially equivalent to optimizing the square loss and softmax loss [23], respectively. More relationships can be found in Table I.

V. A Robust Criterion from Pairwise Loss

Since traditional criteria are related to various pointwise loss, we attempt to derive a new splitting criterion from a pairwise loss, i.e., ranking loss, which has been an important criterion for class-imbalanced learning, cost-sensitive learning, learning to rank, etc. It is defined as

$$l(f; x^+, x^-) = \frac{1}{2}\mathbb{I}[f(x^+) = f(x^-)] + \mathbb{I}[f(x^+) < f(x^-)],$$

where x^+, x^- denotes the positive and negative example.

The total ranking loss of all examples is

$$\begin{aligned} L(f) &= \sum_{x^+ \in S_m^+} \sum_{x^- \in S_m^-} l(f; x^+, x^-) \\ &= \sum_{x^+ \in S^+} \Big(\sum_{x^- \in S^-} l(f; x^+, x^-) + \sum_{x^- \in S_m^- \setminus S^-} l(f; x^+, x^-) \Big) \\ &+ \sum_{x^+ \in S_m^+ \setminus S^+} \Big(\sum_{x^- \in S^-} l(f; x^+, x^-) + \sum_{x^- \in S_m^- \setminus S^-} l(f; x^+, x^-) \Big), \end{aligned}$$

where S_m^+, S_m^- denote the sets of all positive and negative examples respectively, S^+, S^- denote the sets of positive and negative examples in the current node.

Owing to mutual coupling of pairwise loss, we introduce a lower bound to simplify calculation as follows:

$$L(f) \geq \widehat{L}(f) = \sum_{x^+ \in S^+} \sum_{x^- \in S^-} l(f; x^+, x^-).$$

Before splitting, we have

$$\widehat{L}(f) = \frac{1}{2} N_- N_+,$$

since all examples have the same output values.

After splitting, we denote the output values of left and right child by $w_L, w_R \in \mathcal{R}$. Let $N_L^+, N_L^-, N_R^+, N_R^-$ denote the numbers of positive and negative examples in the left and right child, respectively. For $w_L < w_R$, we have

$$\widehat{L}(f) = \widehat{L}(w_L, w_R) = \frac{1}{2} N_L^+ N_L^- + \frac{1}{2} N_R^+ N_R^- + N_L^+ N_R^-.$$

This follows that

$$t(s) = \frac{1}{2} N_- N_+ - \widehat{L}(w_L, w_R) = \frac{1}{2} N_L^+ N_R^- - \frac{1}{2} N_L^- N_R^+.$$

Similarly, we have, for $w_L > w_R$,

$$t(s) = \frac{1}{2} N_- N_+ - \widehat{L}(w_L, w_R) = \frac{1}{2} N_L^- N_R^+ - \frac{1}{2} N_L^+ N_R^-.$$

For $w_L = w_R$, we have

$$t(s) = \frac{1}{2} N_- N_+ - \widehat{L}(w_L, w_R) = 0.$$

Hence, our new splitting criterion, called *pairwise gain*, is defined as

$$\begin{aligned} t(s) &= \frac{1}{2} N_- N_+ - \min_{w_L, w_R} \widehat{L}(w_L, w_R) \\ &= \max_{w_L, w_R} \frac{1}{2} N_- N_+ - \widehat{L}(w_L, w_R) \\ &= \frac{1}{2} |N_L^- N_R^+ - N_L^+ N_R^-|. \end{aligned}$$

We now present a theoretical analysis on the robustness of our proposed criterion *pairwise gain* under label noises.

Theorem 1: Suppose that there are the same label noise proportions τ_+, τ_- in left and right child nodes, noises do not influence the split selection based on *pairwise gain*.

Proof: Let $\widetilde{N}_L^+$, $\widetilde{N}_L^-$, $\widetilde{N}_R^+$ and $\widetilde{N}_R^-$ denote the sizes of positive and negative examples under label noises in left and right child nodes, respectively. We have

$$\begin{aligned} \widetilde{N}_L^+ &= N_L^+(1-\tau_+) + N_L^-\tau_-, \\ \widetilde{N}_L^- &= N_L^-(1-\tau_-) + N_L^+\tau_+, \\ \widetilde{N}_R^+ &= N_R^+(1-\tau_+) + N_R^-\tau_-, \\ \widetilde{N}_R^- &= N_R^-(1-\tau_-) + N_R^+\tau_+. \end{aligned}$$

Then we have

$$\begin{aligned} &\widetilde{N}_L^+ \widetilde{N}_R^- - \widetilde{N}_L^- \widetilde{N}_R^+ \\ =&\left(N_L^+(1-\tau_+) + N_L^-\tau_-\right)\left(N_R^-(1-\tau_-) + N_R^+\tau_+\right) \\ &-\left(N_L^-(1-\tau_-) + N_L^+\tau_+\right)\left(N_R^+(1-\tau_+) + N_R^-\tau_-\right) \\ =&N_L^+N_R^-(1-\tau_+)(1-\tau_-) + N_L^+N_R^+\tau_+(1-\tau_+) \\ &+ N_L^-N_R^-\tau_-(1-\tau_-) + N_L^-N_R^+\tau_+\tau_- \\ &- N_L^-N_R^+(1-\tau_+)(1-\tau_-) - N_L^+N_R^+\tau_+(1-\tau_+) \\ &- N_L^-N_R^-\tau_-(1-\tau_-) - N_L^+N_R^-\tau_+\tau_- \\ =&(1-\tau_+-\tau_-)(N_L^+N_R^- - N_L^-N_R^+). \end{aligned}$$

Table II
BENCHMARK DATASETS

dataset	#instance	#feature	dataset	#instance	#feature
sonar	208	60	pendigits	10992	16
heart	270	13	phishing	11055	68
ionosphere	351	34	letter	20000	16
breast	683	10	protein	24387	357
australian	690	14	a9a	48842	123
diabetes	768	8	shuttle	58000	9
vehicle	846	18	w8a	64700	300
fourclass	862	2	connect4	67557	126
german	1000	24	mnist	70000	780
segment	2310	19	sensit	98528	50
splice	3175	60	ijcnn1	141691	22
dna	3186	180	skin-non	245057	3
satimage	6435	36	webspam	350000	254
gisette	7000	5000	cod-rna	488565	8
mushrooms	8124	112	covtype	581012	54
usps	9298	256	poker	1025010	10

This follows that

$$\begin{aligned} \widetilde{t}(s) &= \frac{1}{2} |\widetilde{N}_L^+ \widetilde{N}_R^- - \widetilde{N}_L^- \widetilde{N}_R^+| \\ &= \frac{1}{2} |1-\tau_+-\tau_-| \cdot |N_L^+ N_R^- - N_L^- N_R^+| \\ &= |1-\tau_+-\tau_-| \cdot t(s). \end{aligned}$$

Hence, we can get

$$\arg\max_s \widetilde{t}(s) = \arg\max_s t(s).$$

Label noises do not influence the split selection based on the *pairwise gain* criterion. This theorem holds. ■

As can be seen, our proposed splitting criterion is robust to both symmetric and asymmetric label noises. Most of the traditional criteria are weighted sums of two impurity measures so that the statistics in left and right nodes are independent. However, they are coupled in our criterion, so the robustness still holds for asymmetric noises.

VI. EXPERIMENTS

The goal is to empirically validate that decision trees and random forests based on our proposed criterion *pairwise gain* can achieve better accuracy under label noises, than ones based on the most frequently used criteria, i.e., gini impurity and information gain, as well as with more compact tree structure and less training time cost.

Parameter Settings. In all experiments, 5-fold cross validation is executed to select stopping parameter, i.e., the minimum number of samples required to split an internal node, denoted by $k \in [2, 10, 40, 80, 150, 500]$. And we set the maximal depth of each tree to be 50 so that a node is forced to be leaf after reaching the maximal tree depth. As for random forests, we randomly choose $sqrt(\#features)$ features when looking for the best split. For each forest, it consists of 100 trees.

Datasets. Thirty-two benchmark datasets[1] are summarized in Table II. Multi-class datasets have been transformed into binary ones by partitioning classes into two groups,

[1]https://www.csie.ntu.edu.tw/˜cjlin/libsvmtools/datasets/

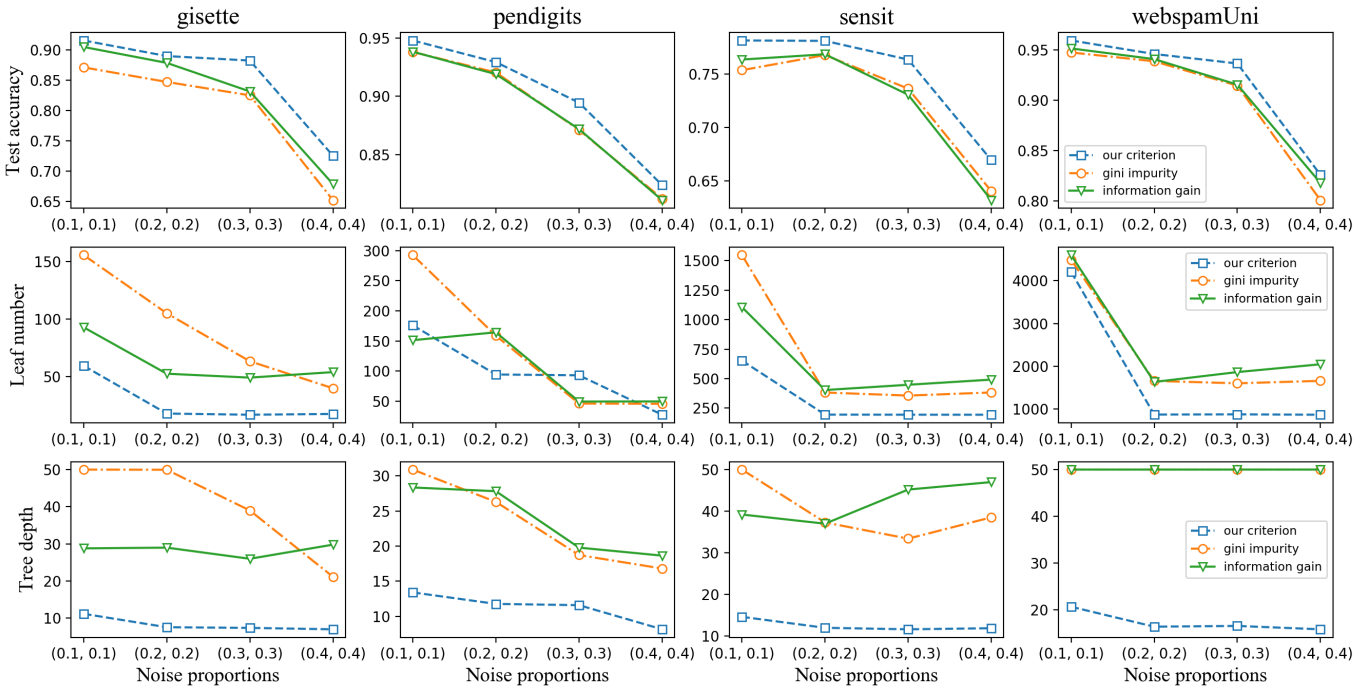

Figure 1. Comparison of test accuracy, leaf number and tree depth under different symmetric noises

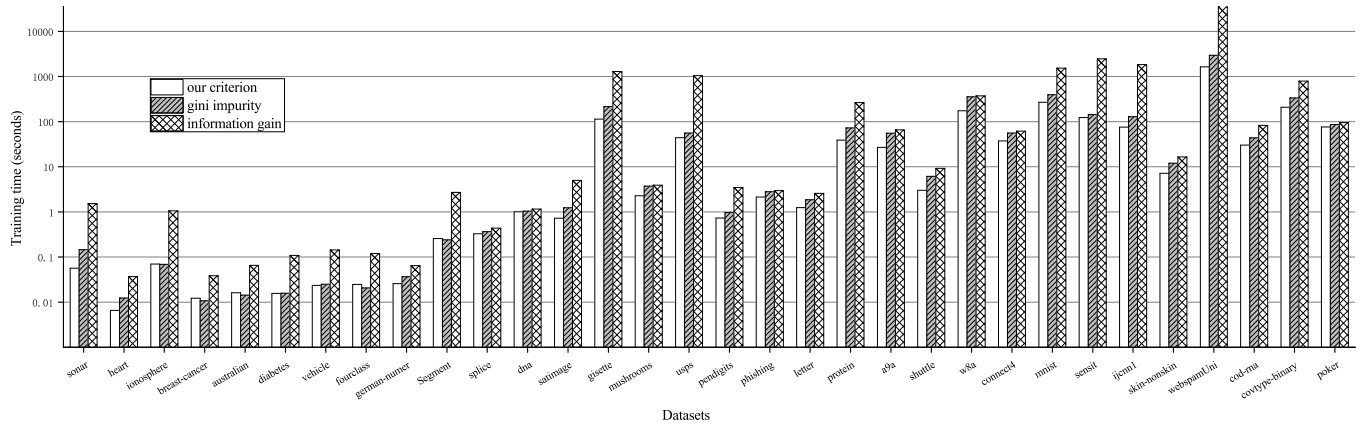

Figure 2. Comparison of the running time (in seconds) on benchmark datasets under noises $(\tau_-, \tau_+) = (0.4, 0.4)$. Notice that the y-axis is in log-scale.

where each group contains the similar sample size. We consider six groups of noise proportions $(\tau_-, \tau_+) \in \{(0.1, 0.1), (0.2, 0.2), (0.3, 0.3), (0.4, 0.4), (0.1, 0.3), (0.2, 0.4)\}$, and labels in the training set are flipped accordingly with different random seeds.

Evaluation metrics. We adopt the test accuracy as the classification performance measurement which is suitable for these balanced datasets. And for decision trees, we further analyze tree structure using leaf number and tree depth. Since a random forest fits many decision trees, we only need to evaluate the efficiency of single decision trees.

All measures of the compared methods are evaluated by 10 trials of 5-fold cross validation with different random seeds, where the performances are obtained by averaging over 50 runs. Experiments are performed using Python on nodes of a computational cluster with 16 CPUs (Intel Xeon Core 3.0GHz) running RedHat with 48GB main memory.

The experimental results are divided into two categories: decision trees under symmetric noises and random forests under asymmetric noises.

A. *Experimental Results for Decision Trees*

Figure 1 show the comparisons of test accuracy, tree structure between our criterion and traditional splitting criteria for decision trees under different symmetric noises $(\tau_-, \tau_+) \in \{(0.1, 0.1), (0.2, 0.2), (0.3, 0.3), (0.4, 0.4)\}$, respectively. We only present the performances on four datasets, while the trends are similar on the other datasets. As can be seen,

Table III

COMPARISON OF TEST ACCURACIES (MEAN±STD.) UNDER LABEL NOISES. ●/○ INDICATES THAT OUR METHOD IS SIGNIFICANTLY BETTER/WORSE THAN THE CORRESPONDING METHODS (PAIRWISE t-TESTS AT 95% SIGNIFICANCE LEVEL).

dataset	(τ_-, τ_+)	*our criterion*	*gini impurity*	*information gain*
sonar	(0.1, 0.3)	0.7456±0.0828	0.7214±0.0822	0.7187±0.0770
	(0.2, 0.4)	0.6856±0.0897	0.6583±0.0889	0.6539±0.0920
heart	(0.1, 0.3)	0.7619±0.0945	0.7352±0.0597	0.7400±0.0661
	(0.2, 0.4)	0.6926±0.0715	0.6626±0.0717●	0.6659±0.0674
ionosphere	(0.1, 0.3)	0.8501±0.0479	0.8452±0.0791	0.8570±0.0755
	(0.2, 0.4)	0.7479±0.0744	0.7400±0.0670	0.7536±0.0641
breast	(0.1, 0.3)	0.9662±0.0127	0.9451±0.0247●	0.9442±0.0249●
	(0.2, 0.4)	0.9633±0.0134	0.9004±0.0502●	0.9009±0.0488●
australian	(0.1, 0.3)	0.8323±0.0368	0.8216±0.0364	0.8246±0.0383
	(0.2, 0.4)	0.7583±0.0527	0.7401±0.0641	0.7361±0.0624
diabetes	(0.1, 0.3)	0.7116±0.0380	0.7070±0.0396	0.7068±0.0413
	(0.2, 0.4)	0.6345±0.0457	0.6294±0.0480	0.6293±0.0547
vehicle	(0.1, 0.3)	0.8929±0.0299	0.8781±0.0348●	0.8805±0.0338
	(0.2, 0.4)	0.8181±0.0383	0.7965±0.0451●	0.7962±0.0422●
fourclass	(0.1, 0.3)	0.9316±0.0325	0.9104±0.0327●	0.9133±0.0305●
	(0.2, 0.4)	0.8211±0.0544	0.7957±0.0568●	0.7920±0.0587●
german	(0.1, 0.3)	0.7140±0.0330	0.7091±0.0413	0.7082±0.0395
	(0.2, 0.4)	0.6205±0.0371	0.5914±0.0510●	0.5895±0.0549●
segment	(0.1, 0.3)	0.9553±0.0128	0.9473±0.0152●	0.9513±0.0165
	(0.2, 0.4)	0.8873±0.0285	0.8774±0.0280	0.8771±0.0284
splice	(0.1, 0.3)	0.9111±0.0165	0.9033±0.0207●	0.8999±0.0201●
	(0.2, 0.4)	0.8184±0.0337	0.7855±0.0442●	0.7788±0.0443●
dna	(0.1, 0.3)	0.8985±0.0136	0.8930±0.0153	0.8883±0.0161●
	(0.2, 0.4)	0.8141±0.0317	0.7809±0.0391●	0.7726±0.0388●
satimage	(0.1, 0.3)	0.9151±0.0084	0.9045±0.0090●	0.9069±0.0094●
	(0.2, 0.4)	0.8674±0.0150	0.8616±0.0156	0.8618±0.0164
gisette	(0.1, 0.3)	0.9308±0.0059	0.9441±0.0051○	0.9464±0.0038○
	(0.2, 0.4)	0.8621±0.0186	0.8794±0.0206	0.8749±0.0226
mushrooms	(0.1, 0.3)	0.9997±0.0005	0.9994±0.0009●	0.9994±0.0008●
	(0.2, 0.4)	0.9901±0.0055	0.9871±0.0069●	0.9870±0.0068●
usps	(0.1, 0.3)	0.9401±0.0047	0.9270±0.0059●	0.9299±0.0068●
	(0.2, 0.4)	0.8746±0.0132	0.8624±0.0140	0.8595±0.0131●
pendigits	(0.1, 0.3)	0.9723±0.0049	0.9696±0.0050●	0.9689±0.0052●
	(0.2, 0.4)	0.9222±0.0116	0.9196±0.0127	0.9154±0.0133●
phishing	(0.1, 0.3)	0.9377±0.0057	0.9341±0.0058●	0.9350±0.0054●
	(0.2, 0.4)	0.9000±0.0112	0.8980±0.0106	0.8977±0.0110
letter	(0.1, 0.3)	0.9091±0.0056	0.8997±0.0062●	0.9006±0.0064●
	(0.2, 0.4)	0.8348±0.0098	0.8272±0.0101●	0.8265±0.0106●
protein	(0.1, 0.3)	0.6407±0.0091	0.6490±0.0102	0.6526±0.0109○
	(0.2, 0.4)	0.5821±0.0117	0.5861±0.0137	0.5842±0.0118
a9a	(0.1, 0.3)	0.8249±0.0037	0.8231±0.0045	0.8233±0.0042
	(0.2, 0.4)	0.7842±0.0049	0.7815±0.0059	0.7804±0.0049
shuttle	(0.1, 0.3)	0.9985±0.0003	0.9986±0.0006	0.9987±0.0005○
	(0.2, 0.4)	0.9976±0.0011	0.9965±0.0013●	0.9966±0.0014●
w8a	(0.1, 0.3)	0.9808±0.0017	0.9848±0.0015○	0.9849±0.0017○
	(0.2, 0.4)	0.9767±0.0011	0.9777±0.0021	0.9773±0.0023
connect4	(0.1, 0.3)	0.8177±0.0026	0.8044±0.0044●	0.8041±0.0040●
	(0.2, 0.4)	0.7348±0.0040	0.7008±0.0083●	0.7022±0.0080●
mnist	(0.1, 0.3)	0.9334±0.0028	0.9269±0.0034●	0.9283±0.0028●
	(0.2, 0.4)	0.8765±0.0055	0.8528±0.0055●	0.8537±0.0066●
sensit	(0.1, 0.3)	0.8041±0.0022	0.8006±0.0025●	0.8033±0.0018
	(0.2, 0.4)	0.7747±0.0031	0.7674±0.0028●	0.7673±0.0033●
ijcnn1	(0.1, 0.3)	0.9815±0.0011	0.9813±0.0007	0.9812±0.0009
	(0.2, 0.4)	0.9612±0.0031	0.9583±0.0028	0.9579±0.0025●
skin-non	(0.1, 0.3)	0.9984±0.0002	0.9982±0.0002●	0.9982±0.0002●
	(0.2, 0.4)	0.9963±0.0006	0.9942±0.0008●	0.9940±0.0009●
webspam	(0.1, 0.3)	0.9689±0.0008	0.9644±0.0011●	0.9662±0.0010●
	(0.2, 0.4)	0.9399±0.0020	0.9358±0.0023●	0.9358±0.0022●
cod-rna	(0.1, 0.3)	0.9589±0.0005	0.9580±0.0004●	0.9585±0.0004
	(0.2, 0.4)	0.9368±0.0006	0.9372±0.0007	0.9375±0.0009
covtype	(0.1, 0.3)	0.8998±0.0023	0.8960±0.0010●	0.8971±0.0009●
	(0.2, 0.4)	0.8333±0.0031	0.8298±0.0018●	0.8293±0.0023●
poker	(0.1, 0.3)	0.7104±0.0059	0.7059±0.0035	0.7055±0.0031●
	(0.2, 0.4)	0.6573±0.0013	0.6462±0.0012●	0.6451±0.0012●
win/tie/loss			**35/27/2**	**35/25/4**

the test accuracy decreases with increasing of symmetric noise proportions, and trees based on our criterion achieve better accuracy than trees based on the gini impurity and information gain. Besides, our criterion leads to more compact tree structure, and the tree depth is significantly smaller than the others. Hence, trees learned using our criterion have less memory cost and more inference efficiency. We can find that the bigger label noises are, the superior our proposed criterion is. And it is natural for decision trees to prefer compact structure in the case of large noise proportions, since the outputs of big trees, whose leaves contain a small number of noisy examples, are probably wrong.

We also compare the training time, and the average CPU time (in seconds) is shown in Figure 2. Our method takes less or comparable running time with the method based on gini impurity. In particular, decision tree induction based on our criterion is about 10 times faster than the method based on information gain, since logarithms are computed in information gain and trees learned by this criterion have much more nodes which need more times of computing splitting criteria during tree induction. As for random forests that fit a few trees, the comparison of time is similar.

B. Experimental Results for Random Forests

Table III show the comparisons of test accuracy between our criterion and traditional splitting criteria for random forests under asymmetric noises $(\tau_-, \tau_+) \in \{(0.1, 0.3), (0.2, 0.4)\}$, respectively. We can observe that random forests based on our criterion achieve better test accuracy than the gini impurity and information gain criteria, even if bagging and random split selection in random forests can improve the robustness to label noises. We think the superiority benefits from the property that our criterion is robust to arbitrary noises, while traditional splitting criteria, like gini impurity and information gain, are not theoretically robust to asymmetric noises. As for symmetric noises, we observe that these criteria get so comparable performances that we do not present the results due to page limitation. We think random forests are inherently robust to symmetric noises no matter which splitting criterion.

VII. Conclusion and Future Work

Splitting criteria have been an important issue on the construction of decision trees. This work presents a unified framework on the splitting criteria from the optimization of loss functions. We point out that decision tree induction based on some classical criteria essentially optimizes different pointwise loss functions, e.g., gini impurity and information gain correspond to the optimization of square loss and softmax loss, respectively. We further derive a new splitting criterion *pairwise gain* from pairwise loss, which is theoretically robust to label noises, including both symmetric and asymmetric noises. Extensive experiments show that decision trees and random forests based on our *pairwise gain* criterion are more robust to label noises, in contrast to the most frequently used criteria, information gain and gini impurity. In the future, an interesting attempt is to exploit more new splitting criteria from other loss functions based on our framework.

Acknowledgment: The research was supported by the National Key R&D Program of China (2017YFB1001903) and NSFC(61876078).

References

[1] L. I. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees (CART)," *Encyclopedia of Ecology*, vol. 40, no. 3, pp. 582–588, 1984.

[2] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[3] W.-Y. Loh and Y.-S. Shih, "Split selection methods for classification trees," *Statistica Sinica*, vol. 7, no. 4, pp. 815–840, 1997.

[4] L. I. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[6] Z.-H. Zhou and J. Feng, "Deep forest: Towards an alternative to deep neural networks," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3553–3559.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[8] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, R. Moore, R. Moore, P. Kohli, A. Criminisi, and A. Kipman, "Efficient human pose estimation from single depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2013.

[9] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, no. 2–3, pp. 81–227, 2012.

[10] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[11] D. Heath, S. Kasif, and S. Salzberg, "Learning oblique decision trees," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1993, pp. 1002–1007.

[12] C. Drummond and R. Holte, "Exploiting the cost (in)sensitivity of decision tree splitting criteria," in *Proceedings of the 17th International Conference on Machine Learning*, vol. 1, no. 1, 2000.

[13] J. Mingers, "An empirical comparison of selection measures for decision tree induction," *Machine learning*, vol. 3, no. 4, pp. 319–342, 1989.

[14] W. Buntine and T. Niblett, "A further comparison of splitting rules for decision-tree induction," *Machine Learning*, vol. 8, no. 1, pp. 75–85, 1992.

[15] W.-Z. Liu and A. P. White, "The importance of attribute selection measures in decision tree induction," *Machine Learning*, vol. 15, no. 1, pp. 25–41, 1994.

[16] U. M. Fayyad and K. B. Irani, "The attribute selection problem in decision tree generation," in *Proceedings of the 10th AAAI Conference on Artificial Intelligence*, 1992, pp. 104–110.

[17] P. C. Taylor and B. W. Silverman, "Block diagrams and splitting criteria for classification trees," *Statistics and Computing*, vol. 3, no. 4, pp. 147–161, 1993.

[18] L. Breiman, "Some properties of splitting criteria," *Machine Learning*, vol. 24, no. 1, pp. 41–47, 1996.

[19] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical statistics*, vol. 15, no. 3, pp. 651–674, 2006.

[20] T. Chen and C. Guestrin, "XGboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.

[21] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *Journal of artificial intelligence research*, vol. 11, pp. 131–167, 1999.

[22] A. Ghosh, N. Manwani, and P. Sastry, "On the robustness of decision tree learning under label noise," in *Proceedings of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2017, pp. 685–697.

[23] A. Painsky and G. Wornell, "On the universality of the logistic loss function," in *Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 936–940.