On the Robustness of Decision Tree Learning Under Label Noise

Aritra Ghosh[1], Naresh Manwani[2], and P.S. Sastry[3(✉)]

[1] Microsoft, Bangalore, India
arghosh@microsoft.com
[2] International Institute of Information Technology, Hyderabad, India
naresh.manwani@iiit.ac.in
[3] Indian Institute of Science, Bangalore, India
sastry@ee.iisc.ernet.in

Abstract. In most practical problems of classifier learning, the training data suffers from label noise. Most theoretical results on robustness to label noise involve either estimation of noise rates or non-convex optimization. Further, none of these results are applicable to standard decision tree learning algorithms. This paper presents some theoretical analysis to show that, under some assumptions, many popular decision tree learning algorithms are inherently robust to label noise. We also present some sample complexity results which provide some bounds on the sample size for the robustness to hold with a high probability. Through extensive simulations we illustrate this robustness.

Keywords: Robust learning · Decision trees · Label noise

1 Introduction

For supervised learning of a classifier, we make use of labeled training data. When the class labels in the training data may be incorrect, it is referred to as label noise. Subjectivity and other errors in human labeling, measurement errors, insufficient feature space are some of the main reasons behind label noise. In many large data problems, labeled samples are often obtained through crowd sourcing and the unreliability of such labels is another reason for label noise. Learning from positive and unlabeled samples can also be cast as a problem of learning under label noise [5]. Thus, learning classifiers in the presence of label noise is an important problem [6].

Decision tree is among the most widely used machine learning approaches [19]. However, not many results are known about the robustness of decision tree learning in presence of label noise. It is observed that label noise in the training data increases size of the learnt tree; detecting and removing noisy examples improves the learnt tree [3]. Through an extensive empirical study it is observed that decision tree learning is fairly robust to label noise [13]. In this paper, we present a theoretical study of robustness of decision tree learning.

J. Kim et al. (Eds.): PAKDD 2017, Part I, LNAI 10234, pp. 685–697, 2017.
DOI: 10.1007/978-3-319-57454-7_53

Most theoretical analyses of learning classifiers under label noise are in the context of risk minimization. The robustness of risk minimization depends on the loss function used. It is proved that any convex potential loss is not robust to uniform or symmetric label noise [9]. While most standard convex loss functions are not robust to symmetric label noise, the 0–1 loss is [11]. A general sufficient condition on the loss function for risk minimization to be robust is derived in [7]. The 0–1 loss, sigmoid loss and ramp loss are shown to satisfy this condition while convex losses such as hinge loss and the logistic loss do not satisfy this condition. Interestingly, we can have a convex loss (which is not a convex potential) that satisfies this sufficient condition and the corresponding risk minimization essentially amounts to a highly regularized SVM [18]. Robust risk minimization strategies under the so called class-conditional (or asymmetric) label noise are also proposed [12,17]. None of these results are applicable for the popular decision tree learning algorithms because they cannot be cast as risk minimization.

In this paper, we analyze learning of decision trees under label noise. We consider some of the popular impurity function based methods for learning of decision trees. We show, in the large sample limit, that under symmetric or uniform label noise the split rule that optimizes the objective function under noisy data is the same as that under noise-free data. We explain how this results in the learning algorithm being robust to label noise (under the large sample limit). We also derive some sample complexity bounds to indicate how large a sample we need at a node. We explain how these results indicate robustness of random forest also. We present empirical results to illustrate this robustness of decision trees and random forests. For comparison we also present results obtained with SVM algorithm.

2 Label Noise and Decision Tree Robustness

In this paper, we only consider binary decision trees for binary classification. We use the same notion of noise tolerance as in [11,18].

2.1 Label Noise

Let $\mathcal{X} \subset \mathcal{R}^d$ be the feature space and let $\mathcal{Y} = \{1, -1\}$ be the class labels. Let $S = \{(\mathbf{x}_1, y_{\mathbf{x}_1}), \ldots, (\mathbf{x}_N, y_{\mathbf{x}_N})\} \in (\mathcal{X} \times \mathcal{Y})^N$ be the *ideal* noise-free data drawn *iid* from a fixed but unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The learning algorithm does not have access to this data. The noisy training data given to the algorithm is $S^\eta = \{(\mathbf{x}_i, \tilde{y}_{\mathbf{x}_i}), i = 1, \cdots, N\}$, where $\tilde{y}_{\mathbf{x}_i} = y_{\mathbf{x}_i}$ with probability $(1 - \eta_{\mathbf{x}_i})$ and $\tilde{y}_{\mathbf{x}_i} = -y_{\mathbf{x}_i}$ with probability $\eta_{\mathbf{x}_i}$. As a notation, for any $\mathbf{x}$, $y_{\mathbf{x}}$ denotes its 'true' label while $\tilde{y}_{\mathbf{x}}$ denotes the noisy label. Thus, $\eta_{\mathbf{x}} = \Pr[y_{\mathbf{x}} \neq \tilde{y}_{\mathbf{x}} \mid \mathbf{x}]$. We use $\mathcal{D}^\eta$ to denote the joint probability distribution of $\mathbf{x}$ and $\tilde{y}_{\mathbf{x}}$.

We say that the noise is *uniform* or *symmetric* if $\eta_{\mathbf{x}} = \eta$, $\forall \mathbf{x}$. Note that, under symmetric noise, a sample having wrong label is independent of the feature vector and the 'true' class of the sample. Noise is said to be *class conditional* or

asymmetric if $\eta_{\mathbf{x}} = \eta_+$, for all patterns of class $+1$ and $\eta_{\mathbf{x}} = \eta_-$, for all patterns of class -1. When noise rate $\eta_{\mathbf{x}}$ is a general function of $\mathbf{x}$, it is termed as *non-uniform* noise. Note that the value of η is unknown to the learning algorithm.

2.2 Criteria for Learning Split Rule at a Node of Decision Trees

Most decision tree learning algorithms grow the tree in top down fashion starting with all training data at the root node. At any node, the algorithm selects a split rule to optimize a criterion and uses that split rule to split the data into the left and right children of this node; then the same process is recursively applied to the children nodes till the node satisfies the criterion to become a leaf. Let $\mathcal{F}$ denote a set of split rules. Suppose, a split rule $f \in \mathcal{F}$ at a node v, sends a fraction a of the samples at v to the left child v_l and the remaining fraction $(1-a)$ to the right child v_r. Then many algorithms select a $f \in \mathcal{F}$ to maximize

$$C(f) = G(v) - (aG(v_l) + (1-a)G(v_r)) \tag{1}$$

where $G(\cdot)$ is a so called impurity measure. There are many such impurity measures. Of the samples at any node v, suppose a fraction p are of positive class and a fraction $q = (1-p)$ are of negative class. Then the Gini impurity is defined by $G_{\text{Gini}} = 2pq$ [1]; entropy based impurity is defined as $G_{\text{Entropy}} = -p\log p - q\log q$ [16]; and misclassification impurity is defined as $G_{\text{MC}} = \min\{p, q\}$. Often the criterion C is called the *gain*. Hence, we also use $\text{gain}_{\text{Gini}}(f)$ to refer to $C(f)$ when G is G_{Gini} and similarly for others.

A split criterion different from impurity is twoing rule [1]. Let p, q, a be as above and let p_l (p_r), q_l (q_r) be the corresponding fractions at the left (right) child v_l (v_r) under split rule f. Then twoing rule selects $f \in \mathcal{F}$ which maximizes $G_{\text{Twoing}}(f) = a(1-a)[|p_l - p_r| + |q_l - q_r|]^2/4$.

2.3 Noise Tolerance of Decision Tree

We want the decision tree learnt with noisy labels to have the same error on noise-free test set as that of the tree learnt using noise-free data. Since label noise is random, on any specific noisey training data, the tree learnt would also be random. Hence, we say the learning method is robust if, in the limit as training set size goes to infinity, the above holds. We now formalize this notion.

Definition 1. *A split criterion C is said to be* noise-tolerant *if*

$$\arg\min_{f\in\mathcal{F}} C(f) = \arg\min_{f\in\mathcal{F}} C^{\eta}(f)$$

where $C(f)$ is the value of the split criterion C for a split rule $f \in \mathcal{F}$ on noise free data and $C^{\eta}(f)$ is the value of the criterion function for f on noisy data, in the limit as the data size goes to infinity.

Let the decision tree learnt from training sample S be represented as $LearnTree(S)$ and let the classification of any $\mathbf{x}$ by this tree be represented as $LearnTree(S)(\mathbf{x})$.

Definition 2. *A decision tree learning algorithm LearnTree is said to be noise-tolerant if*

$$P_{\mathcal{D}}(LearnTree(S)(\mathbf{x}) \neq y_{\mathbf{x}}) = P_{\mathcal{D}}(LearnTree(S^{\eta})(\mathbf{x}) \neq y_{\mathbf{x}})$$

Note that for the above to hold it is sufficient if $LearnTree(S)$ is same as $LearnTree(S^{\eta})$. That is, if the tree learnt with noisy samples is same as that learnt with noise-free samples.[1]

3 Theoretical Results

Robustness of decision tree learning requires the robustness of the split criterion at each non-leaf node and robustness of the labeling rule at each leaf node.

3.1 Robustness of Split Rules

We are interested in comparing, for any specific split rule f, the value of $C(f)$ with its value (in the large sample limit) when there is symmetric label noise.

Let the noise-free samples at a node v be $\{(\mathbf{x}_i, y_i),\ i = 1, \cdots, n\}$. Under label noise, the samples at this node would become $\{(\mathbf{x}_i, \tilde{y}_i),\ i = 1, \cdots, n\}$. Suppose in the noise-free case a split rule f sends n_l of these n samples to the left child, v_l, and $n_r = n - n_l$ to right child, v_r. Since the split rule depends only on the feature vector $\mathbf{x}$ and not the labels, the points that go to v_l and v_r would be the same for the noisy samples also. However, what changes with label noise are the class labels and hence the number of examples of different classes at a node.

Let n^+ and $n^- = n - n^+$ be the number of samples of the two classes at node v in the noise-free case. Let these numbers for v_l and v_r be n_l^+, n_l^- and n_r^+, n_r^-. Let these quantities in the noisy case be denoted by $\tilde{n}^+, \tilde{n}^-, \tilde{n}_l^+, \tilde{n}_l^-$ etc. Define binary random variables, $Z_i,\ i = 1, \cdots, n$, by $Z_i = 1$ iff $\tilde{y}_i \neq y_i$. By definition of symmetric label noise, Z_i are *iid* Bernoulli random variables with expectation η.

Let $p = n^+/n, q = n^-/n = (1-p)$. Let p_l, q_l and p_r, q_r be these fractions for v_l and v_r. Let the corresponding quantities for the noisy case be $\tilde{p}, \tilde{q}, \tilde{p}_l, \tilde{q}_l$ etc. Let $p^{\eta}, q^{\eta}, p_l^{\eta}$ etc. be the values of $\tilde{p}, \tilde{q}, \tilde{p}_l$ in the large sample limit. We have

$$\tilde{p} = \frac{\tilde{n}^+}{n} = \frac{1}{n}\left(\sum_{i:\tilde{y}_i=+1} 1\right) = \frac{1}{n}\left(\sum_{i:y_i=+1} (1 - Z_i) \ + \sum_{i:y_i=-1} Z_i\right) \tag{2}$$

$$\tilde{p}_l = \frac{\tilde{n}_l^+}{n_l} = \frac{1}{n_l}\left(\sum_{i:\mathbf{x}_i\in v_l, \tilde{y}_i=+1} 1\right) = \frac{1}{n_l}\left(\sum_{i:\mathbf{x}_i\in v_l, y_i=+1} (1 - Z_i) \ + \sum_{i:\mathbf{x}_i\in v_l, y_i=-1} Z_i\right)$$

[1] For simplicity, we do not consider pruning of the tree.

All the above expressions involve sums of independent random variables. Hence, by law of large numbers, in the large sample limit we get

$$p^\eta = p(1-\eta) + q\eta = p(1-2\eta) + \eta; \quad p_l^\eta = p_l(1-\eta) + q_l\eta = p_l(1-2\eta) + \eta \quad (3)$$

Note that, under symmetric label noise, $\Pr[Z_i = 1] = \Pr[Z_i = 1|y_i] = \Pr[Z_i = 1|\mathbf{x}_i \in B, y_i] = \eta$, for any subset B of the feature space and this fact is used in deriving Eq. (3).

To find the large sample limit of criterion $C(f)$ under label noise, we need values of the impurity function which in turn needs p^η, q^η, p_l^η etc. which are as given above. For example, the Gini impurity is given by $G(v) = 2pq$ for the noise free case. For the noisy sample, its value can be written as $\tilde{G}(v) = 2\tilde{p}\tilde{q}$. Its value in the large sample limit would be $G^\eta(v) = 2p^\eta q^\eta$. Using the above we can now prove the following theorem about robustness of split criteria.

Theorem 1. *Splitting criterion based on Gini impurity, mis-classification rate and twoing rule are noise-tolerant to symmetric label noise given* $\eta \neq 0.5$.

Proof. We prove robustness of Gini impurity here. Robustness under other criteria can similarly be proved.

Let p and q be the fractions of the two classes at a node v and let a be the fraction of points (under a split rule) at the left child, v_l. Recall that the fraction a is same for noisy and noise-free data. The Gini impurity is $G_{\text{Gini}}(v) = 2pq$. Under symmetric label noise, Gini impurity (under large sample limit) becomes (using Eq. (3)),

$$\begin{aligned} G^\eta_{\text{Gini}}(v) &= 2p^\eta q^\eta = 2[((1-2\eta)p + \eta)((1-2\eta)q + \eta)] \\ &= 2pq(1-2\eta)^2 + (\eta - \eta^2) = G_{\text{Gini}}(v)(1-2\eta)^2 + (\eta - \eta^2) \end{aligned}$$

Similar expressions hold for $G^\eta_{\text{Gini}}(v_l)$ and $G^\eta_{\text{Gini}}(v_r)$. The (large sample) value of criterion or impurity gain of f under label noise can be written as

$$\begin{aligned} &\text{gain}^\eta_{\text{Gini}}(f) = G^\eta_{\text{Gini}}(v) - [a\, G^\eta_{\text{Gini}}(v_l) + (1-a)G^\eta_{\text{Gini}}(v_r)] \\ &= (1-2\eta)^2[G_{\text{Gini}}(v) - a\, G_{\text{Gini}}(v_l) - (1-a)\text{Gini}(v_r)] = (1-2\eta)^2 \text{gain}_{\text{Gini}}(f) \end{aligned}$$

Thus for any $\eta \neq 0.5$, if $\text{gain}_{\text{Gini}}(f^1) > \text{gain}_{\text{Gini}}(f^2)$, then $\text{gain}^\eta_{\text{Gini}}(f^1) > \text{gain}^\eta_{\text{Gini}}(f^2)$. Which means that a maximizer of impurity gain based on Gini index under noise-free samples will be also a maximizer of gain under symmetric label noise, under large sample limit.

Remark: Another popular criterion is impurity gain based on entropy which is not considered in the above theorem. The impurity gain based on entropy is not noise-tolerant as can be shown by a counterexample. Consider a case where a node has n samples (n is large). Suppose, under split rule $f_1(f_2)$ we get $n_l = 0.5n(0.3n)$, $n_l^+ = 0.05n(0.003n)$ and $n_r^+ = 0.25n(0.297n)$. Then it can be easily shown that the best split under no-noise (f_2) does not remain best under 40% noise. However, such counter examples may not be generic and entropy based method may also be robust to label noise in practice.

3.2 Robustness of Labeling Rule at Leaf Nodes

We next consider the robustness of criterion to assign a class label to a leaf node. A popular approach is to take majority vote at the leaf node. We prove that, majority voting is robust to symmetric label noise. We also show that it can be robust to non-uniform noise also under a restrictive condition.

Theorem 2. *Let $\eta_{\mathbf{x}} < 0.5, \forall \mathbf{x}$. (a). Then, majority voting at a leaf node is robust to symmetric label noise. (b). It is also robust to nonuniform label noise if all the points at the leaf node belong to one class in the noise free data.*

Proof. Let p and $q = 1 - p$ be the fraction of positive and negative samples at leaf node v.

(a) Under symmetric label noise, the relevant fractions are $p^\eta = (1-\eta)p+\eta q$ and $q^\eta = (1-\eta)q + \eta p$. Thus, $p^\eta - q^\eta = (1-2\eta)(p-q)$. Since $\eta < 0.5$, $(p^\eta - q^\eta)$ will have the same sign as $(p-q)$, proving robustness of the majority voting.
(b) Let v contain all the points from the positive class. Thus, $p = 1, q = 0$. Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be the samples at v. Under non-uniform noise (with $\eta_{\mathbf{x}} < 0.5, \forall \mathbf{x}$),

$$p^\eta = \frac{1}{n}\sum_{i=1}^{n}(1 - \eta_{\mathbf{x}_i}) > \frac{0.5}{n}\sum_{i=1}^{n} 1 = 0.5 \tag{4}$$

Thus, the majority vote will assign positive label to the leaf node v. This proves the second part of the theorem.

3.3 Robustness of Decision Tree Learning Under Symmetric Label Noise: Large Sample Analysis

We have shown that the split rule that maximizes the criterion function under symmetric label noise is same as the one which maximizes it under noise-free case (under large sample limit). This means, under large sample assumption, the same split rule would be learnt at any node irrespective of whether the labels come from noise-free data or noisy data. (Here we assume for simplicity that there is a unique maximizer of the criterion at each node. Otherwise we need some prefixed rule to break ties. We are assuming that the $\mathbf{x}_i$ at a node are same in the noisy and noise-free cases. These are same at the root. If we learn the same split at the root, then at both its children the samples would be same in the two cases and so on).

Our result for leaf node labeling implies that, under large sample assumption, with majority rule, a leaf node would get the same label under noisy or noise-free data. To conclude that we learn the same tree, we need to examine the rule for deciding when a node becomes a leaf. If this is determined by the depth of the node or number of samples at the node then it is easy to see that the same tree would be learnt with noisy and noise-free data. In many algorithms one makes a node as leaf if no split rule gives positive value to the gain. This will also lead to

learning of the same tree with noisy samples as with noise-free samples, because we showed that the gain under noisy case is a linear function of the gain under noise-free case.

Robustness Against General Noise: In our analysis, we have only considered symmetric label noise. In the case of class-conditional noise, noise rate is same for all feature vectors of a class though it may be different for different classes. In the risk minimization framework, class conditional noise can be handled when the noise rates are known (or can be estimated) [7,12,14,17]. We can extend the analysis presented in Sect. 3.1 to relate expected fraction of examples of a class in the noisy and noise-free cases using the two noise rates. Thus, if the noise rates are assumed known (or can be reliably estimated) it should be possible to extend the analysis here to the case of class-conditional noise. In the general case when noise rates are not known (and cannot be reliably estimated), it appears difficult to establish robustness of impurity based split criteria.

3.4 Sample Complexity Under Noise

We established robustness of decision tree learning algorithms under large sample limit. Hence an interesting question is that of how large the sample size should be for our assertions about robustness to hold with a large probability. We provide some sample complexity bounds in this subsection. (Due to space constraint, we provide proof sketch in Appendix).

Lemma 1. *Let leaf node v have n samples. Under symmetric label noise with $\eta < 0.5$, majority voting will not fail with probability at least $1-\delta$ when $n \geq \frac{2}{\rho^2(1-2\eta)^2}\ln(\frac{1}{\delta})$, where ρ is the difference between fraction of positive and negative samples in the noise-free case.*

The sample size needed increases with increasing noise (η) and decreasing ρ (which can be viewed as 'margin of majority'), which is intuitively clear.

Lemma 2. *Let there be n samples at a non-leaf node v. Given two splits f_1 and f_2, suppose gain (Gini, misclassification, twoing rule) for f_1 is higher than that for f_2. Under symmetric label noise with $\eta \neq 0.5$, gain from f_1 will be higher with probability $1-\delta$ when $n \geq \mathcal{O}(\frac{1}{\rho^2(1-2\eta)^2}\ln(\frac{1}{\delta}))$, where ρ denotes the difference between gain of the two splits in the noise-free case.*

While these results shed some lights on sample complexity, we emphasize that these bounds are loose and are obtained using concentration inequalities. In experimental section, we provide results on how many training samples are needed for robust learning of decision trees on a synthetic dataset.

3.5 Noise Robustness in Random Forest

A random forest [2] is a collection of randomized tree classifiers. We represent the set of trees as $g_n = \{g_n(\mathbf{x}, \pi_1), \cdots, g_n(\mathbf{x}, \pi_m)\}$. Here $\pi_1, \cdots, \pi_m$ are *iid*

random variables, conditioned on data, which are used for partitioning the nodes. Finally, majority vote is taken among the random tree classifiers for prediction. We denote this classifier as $\bar{g}_n$.

In a ***purely random forest classifier***, partitioning does not depend on the class labels. At each step, a node is chosen randomly and a feature is selected randomly for the split. A split threshold is chosen uniformly randomly from the interval of the selected feature. This procedure is done k times. In ***a greedily grown random forest classifier*** each tree is grown greedily by improving impurity with some randomization. At each node, a random subset of features are chosen. Tree is grown by computing the best split among those random features only. Breiman's random forest classifier uses Gini impurity gain [2].

A purely random forest classifier/greedily grown random forest, $\bar{g}_n$, is robust to symmetric label noise with $\eta < 0.5$ under large sample assumption. In purely random forest, randomization is on the partitions and the partitions do not depend on class labels (which may be noisy). We proved robustness of majority vote at leaf nodes under symmetric label noise. Thus, for a purely random forest, the classifier learnt with noisy labels would be same as that learnt with noise-free samples. Similarly for a greedily grown trees with Gini impurity measure, we showed that each tree is robust because of both split rule robustness and majority voting robustness. Thus when large sample assumption holds, greedily grown random forest will also be robust to symmetric label noise. The sample complexity for random forests should be less than that for single decision tree because the ensemble classifier results in some variance reduction. Empirically we observe that, often random forest has better robustness than a single decision tree in finite sample cases.

4 Empirical Illustration

In this section, we illustrate our robustness results for learning of decision trees and random forest. We also present results with SVM whose sensitivity towards noise widely varies [9,11,13,18].

4.1 Dataset Description

We used four 2D synthetic datasets. Details are given below. (Here n denotes total number of samples, p_+, p_- represent the class conditional densities, and $\mathcal{U}(\mathcal{A})$ denotes uniform distribution over set $\mathcal{A}$).

- Dataset 1: Checker board 2×2 Pattern: Data uniform over $[0, 2] \times [0, 2]$ and one class region being $([0, 1] \times [0, 1]) \cup ([1, 2] \times [1, 2])$ and $n = 30000$.
- Dataset 2: Checker board 4×4: Extension of the above to a 4×4 grid.
- Dataset 3: Imbalance Linear Data. $p_+ = \mathcal{U}([0, 0.5] \times [0, 1])$ and $p_- = \mathcal{U}([0.5, 1] \times [0, 1])$. Prior probabilities of classes are 0.9 & 0.1, and $n = 40000$.
- Dataset 4: Imbalance and Asymmetric Linear Data. $p_+ = \mathcal{U}([0, 0.5] \times [0, 1])$ and $p_- = \mathcal{U}([0.5, 0.7] \times [0.4, 0.6])$. Prior probabilities are 0.8 & 0.2, and $n = 40000$.

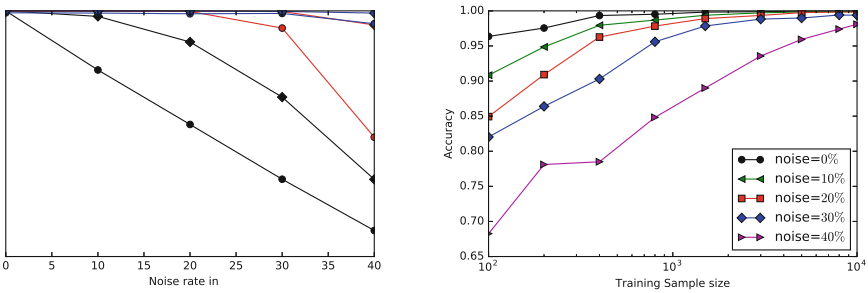
We also present results for 6 UCI datasets [8].

4.2 Experimental Setup

We used decision tree/random forest (RF) implementation in scikit learn library [15]. We present results only with Gini impurity based decision tree classifier. Number of trees in random forest was set to 100. For SVM we used libsvm package [4]. For the results presented in Sect. 4.4, the following setup is used. Minimum leaf size is the only user-chosen parameter in random forest and decision trees. For synthetic datasets, minimum samples in leaf node was restricted to 250. For UCI datasets, it was restricted to 50. For SVM, we used linear kernel (l) for Synthetic Datasets 3, 4 and quadratic kernel (p) for Checker board 2×2 data. In all other datasets we used gaussian kernel (g). For SVM, we selected hyper-parameters using validation data. (Validation range for C is 0.01–500 and for γ in the Gaussian kernel it is 0.001–10). We used 20% data for testing and 20% for validation. Noise rate was varied from 0%–40%. As synthetic datasets are separable, we also experimented with class conditional noise with the two noise rates for the two classes being 40% and 20%. In all experiments, noise was introduced only on training and validation data. Test set was noise free.

4.3 Effect of Sample Size on Robustness of Learning

Here we present experimental results on the test accuracy for different sample sizes using the 2×2 checker board data. We choose a leaf sample size and learn decision tree and random forest with different noise levels. (The training set size is fixed at 20000). We do this for a number of choices for leaf sample size. The test accuracies in all these cases are shown in Fig. 1(a). As can be seen from the figure, even when training data size is huge, we do not get robustness if leaf sample size is small. This is in accordance with our analysis (as in Lemma 1) because minimum sample size is needed for the majority rule to be correct with a large probability. A leaf sample size of 50 seems sufficient to take care of even 30% noise.



(a) (b)

Fig. 1. For 2×2 Checker board data variation of accuracy with (a) Minimum leaf size, (b) Training data size, for different noise levels for DT

Next we experiment with varying the (noisy) training data size. The results are shown in Fig. 1(b). It can be seen that with 400/4000 sample size decision tree learnt has good test accuracy (95%) at 20%/40% noise (the sample ratio is close to $\frac{(1-2\times0.4)^2}{(1-2\times0.2)^2} = 1/9$ as provided in Lemma 1). We need larger sample size for higher level of noise. This is also as expected from our analysis.

4.4 Comparison of Accuracies of Learnt Classifiers

The average test accuracy and standard deviation (over 10 runs) on different data sets under different levels of noise are shown in Table 1 for synthetic datasets and in Table 2 for UCI datasets. In Table 2 we also indicate the dimension of feature vector (d), the number of positive and negative samples in the data (n^+, n^-).

For synthetic datasets, the sample sizes are large and hence we expect good robustness. As can be seen from Table 1, for noise-free data, all classifiers (decision tree, random forest and SVM) perform equally. However, with 30% or 40% noise, the accuracies of SVM are much poorer than those of decision tree and random forest. For example, for synthetic datasets 3 and 4, the average accuracies of decision tree and random forest classifiers continue to be 99% even at 40% noise while those of SVM drop to about 90% and 80% respectively. Note that even with very large sample sizes, we do not get robustness in SVM. It can be seen that decision tree and random forest classifiers are robust to class conditional noise also, even without knowledge about noise rate (as indicated by last column in the table). Our current analysis does not prove this robustness; this is one possible extension of the theoretical analysis presented here.

Table 1. Comparison of accuracies on synthetic datasets

Data	Method	$\eta = 0\%$	$\eta = 10\%$	$\eta = 20\%$	$\eta = 30\%$	$\eta = 40\%$	$\eta_+ = 40\%$ $\eta_- = 20\%$
2 × 2 CB	Gini	99.95 ± 0.05	99.9 ± 0.06	99.91 ± 0.1	99.82 ± 0.16	98.97 ± 0.83	99.45 ± 0.83
	RF	99.99 ± 0.02	99.96 ± 0.02	99.91 ± 0.05	99.87 ± 0.06	99.16 ± 0.18	99.11 ± 0.45
	SVM (p)	99.83 ± 0.12	97.38 ±1.21	91.88 ± 2.65	87.96 ± 5.52	76.42 ± 4.43	68.78 ± 0.97
4 × 4 CB	Gini	99.76 ± 0.18	99.72 ± 0.16	99.46 ± 0.18	98.71 ± 0.32	95.21 ± 1.08	97.36 ± 1.23
	RF	99.94 ± 0.02	99.9 ± 0.02	99.78 ± 0.04	99.35 ± 0.15	96.23 ± 0.91	95.41 ± 0.53
	SVM (g)	99.6 ± 0.05	98.58 ± 0.23	97.81 ± 0.24	96.83 ± 0.46	92.22 ± 2.5	91.24 ± 0.85
Dataset 3	Gini	100.0 ± 0.01	100.0 ± 0.01	99.99 ± 0.01	99.99 ± 0.02	99.92 ± 0.07	99.92 ± 0.18
	RF	100.0 ± 0.01	100.0 ± 0.01	99.99 ± 0.01	99.98 ± 0.02	99.86 ± 0.12	99.9 ± 0.13
	SVM (l)	99.89 ± 0.04	96.65 ± 0.26	90.02 ± 0.3	90.02 ± 0.3	90.02 ± 0.3	90.1 ± 0.31
Dataset 4	Gini	100.0 ± 0.0	99.99 ± 0.01	99.99 ± 0.01	99.98 ± 0.03	99.73 ± 0.54	99.88 ± 0.26
	RF	100.0 ± 0.0	99.99 ± 0.01	99.99 ± 0.01	99.93 ± 0.09	99.91 ± 0.11	99.7 ± 0.31
	SVM (l)	99.86 ± 0.03	99.21 ± 0.24	96.55 ± 4.05	79.96 ± 0.34	79.96 ± 0.34	79.96 ± 0.34

Similar performance is seen on UCI datasets also as shown in Table 2. For breast cancer dataset, there is a small drop in the average accuracy of decision tree with increasing noise rate while for random forest the drop is significantly less. This is also expected because the total sample size here is less. Although

SVM has significantly higher average accuracy than decision tree in 0% noise, at 40% noise its average accuracy drops more than that of decision tree. In all other data sets also, decision tree and random forest are more robust than SVM as can be seen from the table.

Table 2. Comparison of accuracies on UCI datasets

Data (d, n^+, n^-)	Method	$\eta = 0\%$	$\eta = 10\%$	$\eta = 20\%$	$\eta = 30\%$	$\eta = 40\%$
Breast cancer $(10, 239, 444)$	Gini	92.04 ± 3.0	90.36 ± 3.02	90.0 ± 2.24	90.22 ± 2.38	87.23 ± 7.72
	RF	96.64 ± 0.93	96.79 ± 1.23	96.64 ± 1.82	95.91 ± 1.47	96.13 ± 1.39
	SVM	96.79 ± 1.67	96.06 ± 1.91	95.91 ± 2.27	93.72 ± 4.55	92.48 ± 3.62
German $(24, 300, 700)$	Gini	71.2 ± 3.47	71.7 ± 2.5	71.25 ± 3.16	70.25 ± 2.75	64.65 ± 6.29
	RF	70.75 ± 2.71	70.8 ± 2.94	70.9 ± 2.84	71.05 ± 2.44	69.35 ± 3.41
	SVM	75.25 ± 5.45	74.45 ± 3.68	72.1 ± 2.37	69.45 ± 3.06	64.55 ± 7.18
Splice $(60, 1648, 1527)$	Gini	91.26 ± 1.65	91.23 ± 1.61	90.22 ± 1.53	86.22 ± 4.11	74.38 ± 5.54
	RF	94.76 ± 0.68	93.94 ± 0.76	93.87 ± 1.39	91.97 ± 1.82	82.69 ± 3.05
	SVM	91.1 ± 0.77	88.83 ± 1.08	87.67 ± 1.09	83.04 ± 1.36	70.47 ± 6.58
Spam $(57, 1813, 2788)$	Gini	89.74 ± 1.15	89.01 ± 1.86	87.61 ± 2.05	84.57 ± 1.83	80.8 ± 3.0
	RF	92.07 ± 1.1	92.2 ± 0.91	92.06 ± 1.15	91.04 ± 1.95	88.81 ± 1.5
	SVM	89.2 ± 1.02	86.41 ± 0.88	82.55 ± 1.72	76.64 ± 2.28	68.02 ± 3.95
Wine (white) $(11, 3258, 1640)$	Gini	75.36 ± 0.76	74.72 ± 1.69	73.56 ± 1.34	73.08 ± 1.94	69.4 ± 5.72
	RF	76.4 ± 1.38	76.74 ± 1.22	76.45 ± 1.18	74.74 ± 3.27	72.89 ± 1.89
	SVM	75.34 ± 0.76	72.43 ± 1.73	71.08 ± 2.0	68.07 ± 2.18	65.24 ± 2.71
Magic $(10, 12332, 6688)$	Gini	83.75 ± 0.42	83.58 ± 0.49	82.33 ± 0.56	81.36 ± 1.08	78.0 ± 1.74
	RF	85.24 ± 0.58	85.37 ± 0.61	85.3 ± 0.58	84.83 ± 0.71	82.37 ± 1.34
	SVM	82.7 ± 0.43	82.24 ± 0.45	81.0 ± 0.34	79.16 ± 0.43	69.5 ± 3.33

5 Conclusion

In this paper, we investigated the robustness of decision tree learning under label noise. We proved that decision tree algorithms based on Gini or misclassification impurity and the twoing rule algorithm are all robust to symmetric label noise. We also provided some sample complexity results for the robustness. Through empirical investigations we illustrated the robust learning of decision tree and random forest. Decision tree approach is very popular in many practical applications. Hence, the robustness results presented in this paper are interesting. Though we considered only impurity based methods, there are other algorithms for learning decision trees (e.g., [10]). Extending such robustness results to other decision tree learning algorithms is an interesting problem. All the results we proved are for symmetric noise. Extending these results to class conditional and non-uniform noise is another important direction for future research.

A Proof Sketch of Lemmas 1, 2

Let $n^+(\tilde{n}^+)$ and $n^-(\tilde{n}^-)$ denote the positive and negative samples at the node under noise-free case (noisy case). Taking positive class as majority, we note

$\rho = (n^+ - n^-)/n$. Using Hoeffding bound it is easy to show $\Pr[\tilde{n}^+ - \tilde{n}^- < 0] \leq \exp\left(-\frac{\rho^2 n(1-2\eta)^2}{2}\right)$. This gives bound for samples needed as $n > \frac{2}{\rho^2(1-2\eta)^2}\ln(\frac{1}{\delta})$, completing proof of Lemma 1.

Let n, n_l, n_r be the number of samples at v, v_l, v_r and recall $n_l = an$ and $n_r = (1-a)n$. Recall that $\tilde{p}, \tilde{p}_l, \tilde{p}_r$ are fraction of positive samples at v, v_l, v_r and $p^\eta, p_l^\eta, p_r^\eta$ are their large sample values. Then, using Hoeffding bounds we get (with $\epsilon_1 = \epsilon$, $\epsilon_2 = \epsilon/\sqrt{a}$ and $\epsilon_3 = \epsilon/\sqrt{1-a}$),

$$\Pr\left[\left(|\tilde{p} - p^\eta| \geq \epsilon_1\right) \cup \left(|\tilde{p}_l - p_l^\eta| \geq \epsilon_2\right) \cup \left(|\tilde{p}_r - p_r^\eta| \geq \epsilon_3\right)\right] \leq 6e^{-2n\epsilon^2} \tag{5}$$

When this event happens, with some algebraic manipulation, one can show for Gini impurity, $|\hat{\text{gain}}^\eta_{\text{Gini}}(f) - \text{gain}^\eta_{\text{Gini}}(f)| \leq 6(1-2\eta)\epsilon$ where $\hat{\text{gain}}^\eta_{\text{Gini}}$ is the random Gini-gain under noise with sample size n and $\text{gain}^\eta_{\text{Gini}}$ is its large sample limit. This gives us the bound as needed in Lemma 2. We can prove the lemma for other criteria also similarly.

References

1. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, Monterey (1984)
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
3. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. J. Artif. Intell. Res. **11**, 131–167 (1999)
4. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27 (2011)
5. du Plessis, M.C., Niu, G, Sugiyama, M.: Analysis of learning from positive and unlabeled data. In: Advances in Neural Information Processing Systems (2014)
6. Frénay, B., Verleysen, M.: Classification in the presence of label noise: a survey. IEEE Trans. Neural Netw. Learn. Syst. **25**, 845–869 (2014)
7. Ghosh, A., Manwani, N., Sastry, P.S.: Making risk minimization tolerant to label noise. Neurocomputing **160**, 93–107 (2015)
8. Lichman, M.: UCI machine learning repository (2013)
9. Long, P.M., Servedio, R.A.: Random classification noise defeats all convex potential boosters. Mach. Learn. **78**(3), 287–304 (2010)
10. Manwani, N., Sastry, P.S.: Geometric decision tree. IEEE Trans. Syst. Man Cybern. **42**(1), 181–192 (2012)
11. Manwani, N., Sastry, P.S.: Noise tolerance under risk minimization. IEEE Trans. Cybern. **43**(3), 1146–1151 (2013)
12. Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. In: Advances in Neural Information Processing Systems (2013)
13. Nettleton, D.F., Orriols-Puig, A., Fornells, A.: A study of the effect of different types of noise on the precision of supervised learning techniques. Artif. Intell. Rev. **33**(4), 275–306 (2010)
14. Patrini, G., Nielsen, F., Nock, R., Carioni, M.: Loss factorization, weakly supervised learning and label noise robustness. In: Proceedings of The 33rd International Conference on Machine Learning, pp. 708–717 (2016)
15. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

16. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
17. Scott, C., Blanchard, G., Handy, G.: Classification with asymmetric label noise: consistency and maximal denoising. In: The 26th Annual Conference on Learning Theory, 12–14 June 2013, pp. 489–511 (2013)
18. van Rooyen, B., Menon, A., Williamson, R.C.: Learning with symmetric label noise: the importance of being unhinged. In: Advances in Neural Information Processing Systems, pp. 10–18 (2015)
19. Wu, X., et al.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37 (2007)