Alternating Decision Forests

Yufeng Jiang

1. Introduction

Interestingly and besides their simplicity, it is yet not fully theoretically understood what makes RFs such a powerful learning method. Existing explanations in the literature point to comparisons with nearest neighbor algorithms, rules of large numbers [2], classifier consistency [1], and large-margin methods, among others.

Besides these definitely valid insights, authors argue that RFs share one important characteristic with other powerful classifiers like SVMs or Boosting. In contrast to other methods, RFs minimize this loss greedily and implicitly via recursively reducing the uncertainty of given training samples by using independent base classifiers, *i.e.* trees. Although these characteristics result in both, fast and parallel training capabilities, there is no control over an overall classifier loss and its proper minimization. While this makes it theoretically hard and somewhat unintuitive to comprehend the success of this learning method, it also unveils several practical disadvantages. First, during training there is no guarantee that all parameters have been learned properly by the entire model. Second, unnecessary emphasis is given on easy to classify training samples, often leading to too complex models. Third, it is hard to extend the learner to special learning tasks, such as, domain adaptation, semi-supervised or multiple instance learning, as this is usually realized via regularizing a global loss function.

2. Introducing a Global Loss

To allow for integrating different, global loss functions into the Random Forests training procedure, they adopt ideas from Boosting [3]. A Boosting classifier $F(x)$ consists of T weak learners $f_t(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{R}^K$, where each weak learner gives a prediction $p_t(y|\mathbf{x})$ about the calss confidences for a sample $\mathbf{x}$. The final output of a Boosting classifier is the weighted sum of the class confidences, such as, $F(x) = \sum_{t=1}^{T} v_t f_t(\mathbf{x})$, where v_t steers the influence of each individual weak learner.

Friedman *et al.* [4] showed that Boosting can be understood as performing Gradient Descent in function space, paving the way for a new Boosting method named $GradientBoost$. In more detail, given a labeled training set $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$, training a single weak learner can be written as the global loss minimization problem

$$\arg\min_{\Theta^t} \sum_{i=1}^{N} l(y_i; \mathcal{F}_{1:t-1}(\mathbf{x}_i; \bar{\Theta}) + f_t(\mathbf{x}_i; \Theta^t)). \quad (1)$$

Here, $l(\cdot)$ is a differentiable loss function, $\mathcal{F}_{1:t-1}(\mathbf{x}; \bar{\Theta}) = \sum_{j-1}^{t-1} v \cdot f_j(\mathbf{x}; \Theta^j)$ describes the already trained classifier and $f_t(\mathbf{x}; \Theta^t)$ is the classifier in the current iteration t; $\bar{\Theta}$ is the collection of the parameters of the already fixed weak learners and Θ^t are the parameters to be trained in the current iteration; v is the so-called shrinkage factor.

With a Taylor expansion, they can re-write Eq.1 as

$$\arg\min_{\Theta^t} \sum_{i=1}^{N} l(y_i; \mathcal{F}_{1:t-1}(\mathbf{x}_i; \bar{\Theta})) - \frac{\partial l(y_i, \mathcal{F}_{1:t-1}(\mathbf{x}_i; \bar{\Theta}))}{\partial \mathcal{F}(\mathbf{x})} \cdot f_t(\mathbf{x}_i; \Theta^t), \quad (2)$$

in order to learn the parameters Θ^t of the current weak classifier. This can be done by training $f_t(\mathbf{x}; \Theta^t)$ to have high correlation with the negative gradient of the loss, which corresponds to updating the weights w_i^t for each training sample $\mathbf{x}_i$ in iteration t as

$$w_i^t = \left| \frac{\partial l(y_i, \mathcal{F}_{1:t-1}(\mathbf{x}_i; \bar{\Theta}))}{\partial \mathcal{F}(\mathbf{x})} \right| . \quad (3)$$

This process of $alternating$ between training a single stage d and updating the weights w_i^{d+1} for the next stage is repeated until the same stopping criteria as in standard RFs are reached. Hence, authors name this learning method $Alternating\ Decidion\ Forests$. They give an illustrative overview of this scheme in Fig.1. Futhermore, they note that inference in ADFs is exactly the same as in RFs, *i.e.* ADFs also inherit the properties of low computational costs during the testing phase from RFs.

3. Machine Learning Experiments

Their experiments on machine learning benchmarks give a detailed analysis of the proposed classifier. First, they compare ADFs with the most related competing methods,

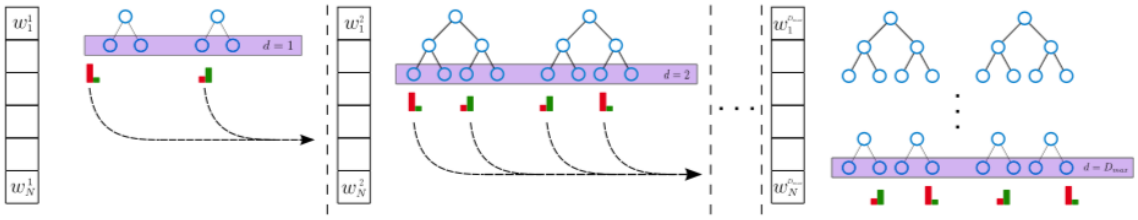

Figure 1. Overview of the proposed tree growing principle of *Alternating Decision Forests*. In the first iteration (d = 1), the weights w_i^d are uniform, and the first split functions are trained in a breadth-first manner. This forest with depth d = 2 can give predictions on the training samples, which are used to calculate weights, based on a global loss function, for the next iteration d = 3. This procedure is repeated until the maximum tree depth $d = D_{max}$ is reached.

Dataset	# Train	# Test	# Features	# Classes
$G50c$	50	500	50	2
$Letter$	16000	4000	16	26
$USPS$	7291	2007	256	10
$MINIST$	60000	10000	784	10
$Char74k$	66707	7400	64	62

Table 1. Properties of the machine learning data sets used in our evaluation.

i.e. Random Forests (RFs) and Boosted Trees (BTs) on 5 different data sets. They also investigate different choices of the loss function. Then, they evaluate the influence of two important parameters common to ADFs, RFs and BTs on the overall classification performance. Finally, they give a dense evaluation of different parameter choices for their classifier.

Data sets: They use 5 standard machine learning data sets to compare ADFs with related approaches and also to investigate different paremeter settings. The properties of these data sets are summarized in Tab.1.

References

[1] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *JMLR*, 9(9):2015–2033, 2008. 1

[2] L. Breiman. Random forests. *ML*, 45(1):5–32, 2001. 1

[3] Y. Freund and R. E. Shapire. Experiments with a new boosting algorithm. In *ICML*, 1996. 1

[4] J. J. Hull. A database for handwritten text recognition research. *IEEE TPAMI*, 16(5):550–554, 1994. 1