Pushing the Envelope of Gradient Boosting Forests via Globally-Optimized Oblique Trees

Magzhan Gabidolla and Miguel Á. Carreira-Perpiñán

Dept. of Computer Science & Engineering
University of California, Merced

CVPR 2022
Gradient Boosting (GB) Forests

- Ensembles of decision trees have long been established as some of the most powerful, off-the-shelf machine learning models.
- In recent years, one type of forest, Gradient Boosting (GB), has gained prominence due to their:
  - Strong empirical performance on many problems
  - The development of extremely efficient implementations such as XGBoost or LightGBM.
- They typically require little effort on hyperparameter tuning and are thus considered “off-the-shelf”.
- Given the tremendous effort put on the development and refinement of the popular GB toolkits, how can we further improve GB forests?
Modeling high-order feature interactions: Axis-aligned vs Oblique trees

- Only 5 features participate in the routing function of the above leaf.
- Max order of feature interactions is limited by the depth $\Delta$ in axis-aligned trees.

- Each decision node is a function of all the features.
- Their non-linear combination is a much more complex order-$D$ interaction.
- For modeling complex functions, a forest of oblique trees should achieve higher accuracy and require fewer and shallower trees.
Synthetic MNIST binary classification

Imagine splitting $28 \times 28$ pixel image into 4 quadrants $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$. Let $Q_i$ be the sum of $[0, 1]$ pixel intensities in quadrant $i$, and $P = 30$.

$$(Q_1 + Q_4) - (Q_2 + Q_3) \geq P$$

$$(Q_2 + Q_3) - (Q_1 + Q_4) \geq P$$

class 1

$E_{\text{test}}(\%)$ vs. Number of trees $T$ in XGBoost
Tree Alternating Optimization (TAO) for GB objective function

Let $\tau(x; \Theta)$ be a binary decision tree of some predetermined structure with parameters $\Theta = \{(w_i, w_{i0})\}_{i \in D} \cup \{\theta_i\}_{i \in L}$, decision nodes in set $D$ and leaves in set $L$.

$$\min_{\Theta} \sum_{n=1}^{N} l(g_n, H_n, \tau(x_n; \Theta)) + \alpha \sum_{i \in D} \|w_i\|_1$$

where $$l(g, H, \gamma) = g^T \gamma + \frac{1}{2} \gamma^T H \gamma.$$
Alternating optimization and separability condition

- Any set of non-descendant nodes of a tree can be optimized independently:

\( R_i \)-reduced set

Fixed
Reduced problem over a decision node

- Evaluate loss induced by left/right subtrees;
- Generate pseudolabel for each instance in reduced set $\mathcal{R}_i$;
- Solve weighted binary classification problem (linear):
The reduced problem takes the form:

$$\min_{w_i, w_i0} \sum_{n \in \mathcal{R}_i} \bar{L}(g_n, H_n, f_i(x; w_i, w_i0)) + \alpha \|w_i\|_1. \quad (1)$$

This problem is NP-hard but can be well approximated with a convex surrogate; we use $\ell_1$-regularized logistic regression, and solve it using LIBLINEAR [1].

The reduced problem consists of optimizing the original loss but over the leaf classifier on its reduced set:

$$\min_{\theta_i} \sum_{n \in \mathcal{R}_i} g_n^T \theta_i + \frac{1}{2} \theta_i^T H_n \theta_i. \quad (2)$$

If $\sum_{n \in \mathcal{R}_i} H_n$ is positive definite, the exact solution is $\theta_i = -\left(\sum_{n \in \mathcal{R}_i} H_n\right)^{-1} \sum_{n \in \mathcal{R}_i} g_n$. In practice either $\theta_i$ is scalar (e.g. binary classification) or one uses a diagonal approximation to the Hessian.
Experimental results: comparison

**News20**

- **GB-sklearn**
- **XGBoost**
- **LightGBM**
- **GB-TAO**

**pendigits**

- **GB-sklearn**
- **XGBoost**
- **LightGBM**
- **GB-TAO**

**News20**

- **XGBoost**
- **LightGBM**
- **GB-TAO**
- **GB-sklearn**

**CT-slice**

- **XGBoost**
- **LightGBM**
- **GB-TAO**
- **GB-sklearn**
Conclusion

• We have motivated the use of a significantly more powerful tree type having hyperplane splits, which are able to learn many-feature interactions effectively.

• Key to this is the ability to optimize the GB loss over such trees, a difficult problem which we address using a variation of tree alternating optimization.

• In raw accuracy, the oblique forests consistently improve over all competitors, sometimes by a surprisingly large margin, using few, shallow trees, often having fewer parameters overall.

• Our work also suggests that exploring other types of trees or loss functions, properly optimized, may result in even better GB forests.

• Work supported by NSF award IIS–2007147
References