Supplementary Material: Feature-Budgeted Random Forest

Feng Nan FNAN@BU.EDU

Joseph Wang JOEWANG@BU.EDU

Venkatesh Saligrama SRV@BU.EDU

Boston University, 8 Saint Mary's Street, Boston, MA

Proof of Lemma 2.3 Before showing admissibility of the threshold-Pairs function in the multiclass setting, we first show $F_\alpha(G)$ is admissible for the binary setting. Consider the binary classification setting, let

$$F_\alpha(G) = [[n_G^1 - \alpha]_+[n_G^2 - \alpha]_+ - \alpha^2]_+.$$

All the properties are obviously true except supermodularity. To show supermodularity, suppose $R \subseteq G$ and object $j \notin R$. Suppose j belongs to the first class. We need to show

$$F_\alpha(G \cup j) - F_\alpha(G) \geq F_\alpha(R \cup j) - F_\alpha(R). \tag{1}$$

Consider 3 cases:
(1) $F_\alpha(R) = F_\alpha(R \cup j) = 0$: The right hand side of (1) is 0 and (1) holds because of monotonicity of F_α.
(2) $F_\alpha(R) = 0, F_\alpha(R \cup j) > 0, F_\alpha(G) = 0$: (1) reduces to $F_\alpha(G \cup j) \geq F_\alpha(R \cup j)$, which is true by monotonicity.
(3) $F_\alpha(R) = 0, F_\alpha(R \cup j) > 0, F_\alpha(G) > 0$: Note that $F_\alpha(G) > 0$ implies that $[n_G^1 - \alpha]_+[n_G^2 - \alpha]_+ - \alpha^2 > 0$ which further implies $n_G^1 > \alpha, n_G^2 > \alpha$. Thus the left hand side is

$$\begin{aligned} &F_\alpha(G \cup j) - F_\alpha(G) = \\ &(n_G^1 - \alpha + 1)(n_G^2 - \alpha) - \alpha^2 - ((n_G^1 - \alpha)(n_G^2 - \alpha) - \alpha^2) \\ &= n_G^2 - \alpha. \end{aligned}$$

The right hand side is

$$\begin{aligned} F_\alpha(R \cup j) &= (n_R^1 - \alpha + 1)(n_R^2 - \alpha) - \alpha^2 \\ &= (n_R^1 - \alpha)(n_R^2 - \alpha) - \alpha^2 + (n_R^2 - \alpha). \end{aligned}$$

If $n_R^1 \geq \alpha$, $F_\alpha(R) = \max((n_R^1 - \alpha)(n_R^2 - \alpha) - \alpha^2, 0) = 0$ because $F_\alpha(R \cup j) > 0$ implies $n_R^2 > \alpha$. So $F_\alpha(R \cup j) \leq n_R^2 - \alpha \leq n_G^2 - \alpha = F_\alpha(G \cup j) - F_\alpha(G)$.
(4) $F_\alpha(R) > 0$: We have

$$F_\alpha(G \cup j) - F_\alpha(G) = n_G^2 - \alpha \geq n_R^2 - \alpha = F_\alpha(R \cup j) - F_\alpha(R).$$

This completes the proof for the binary classification setting. To generalize to the multiclass threshold-Pairs function, again, all properties are obviously true except supermodularity, which follows from the fact that each term in the sum is supermodular according to the proof for binary setting.

More Admissible Impurity Functions The following polynomial impurity function is also admissible.

Lemma 0.1. *Suppose there are k classes in G. Any polynomial function of $n_G^1, \ldots, n_G^k$ with non-negative terms such that $n_G^1, \ldots, n_G^k$ do not appear as singleton terms is* admissible. *Formally, if*

$$F(G) = \sum_{i=1}^{M} \gamma_i (n_G^1)^{p_{i1}} (n_G^2)^{p_{i2}} \ldots (n_G^k)^{p_{ik}}, \tag{2}$$

where γ_i's are non-negative, p_{ij}'s are non-negative integers and for each i there exists at least 2 non-zero p_{ij}'s, then F is admissible.

Proof. Properties (1),(2),(3) and (5) are obviously true. To show F is supermodular, suppose $R \subset G$ and object $\hat{j} \notin R$ and $\hat{j}$ belongs to class j, we have

$$\begin{aligned}
&F(R \cup \hat{j}) - F(R) \\
&= \sum_{i \in I_j} \gamma_i [(n_R^1)^{p_{i1}} \dots (n_R^j + 1)^{p_{ij}} \dots (n_R^k)^{p_{ik}} - \\
&\quad (n_R^1)^{p_{i1}} \dots (n_R^j)^{p_{ij}} \dots (n_R^k)^{p_{ik}}] \\
&\leq \sum_{i \in I_j} \gamma_i [(n_G^1)^{p_{i1}} \dots (n_G^j + 1)^{p_{ij}} \dots (n_G^k)^{p_{ik}} - \\
&\quad (n_G^1)^{p_{i1}} \dots (n_G^j)^{p_{ij}} \dots (n_G^k)^{p_{ik}}] \\
&= F(G \cup \hat{j}) - F(G),
\end{aligned}$$

where the first summation index set I_j is the set of terms that involve n_R^j. The inequality follows because $(n_R^j + 1)^{p_{ij}}$ can be expanded so the negative term can be canceled, leaving a sum-of-products form for R, which is term-by-term dominated by that of G. □

Another family of *admissible* impurity functions is the Powers function.

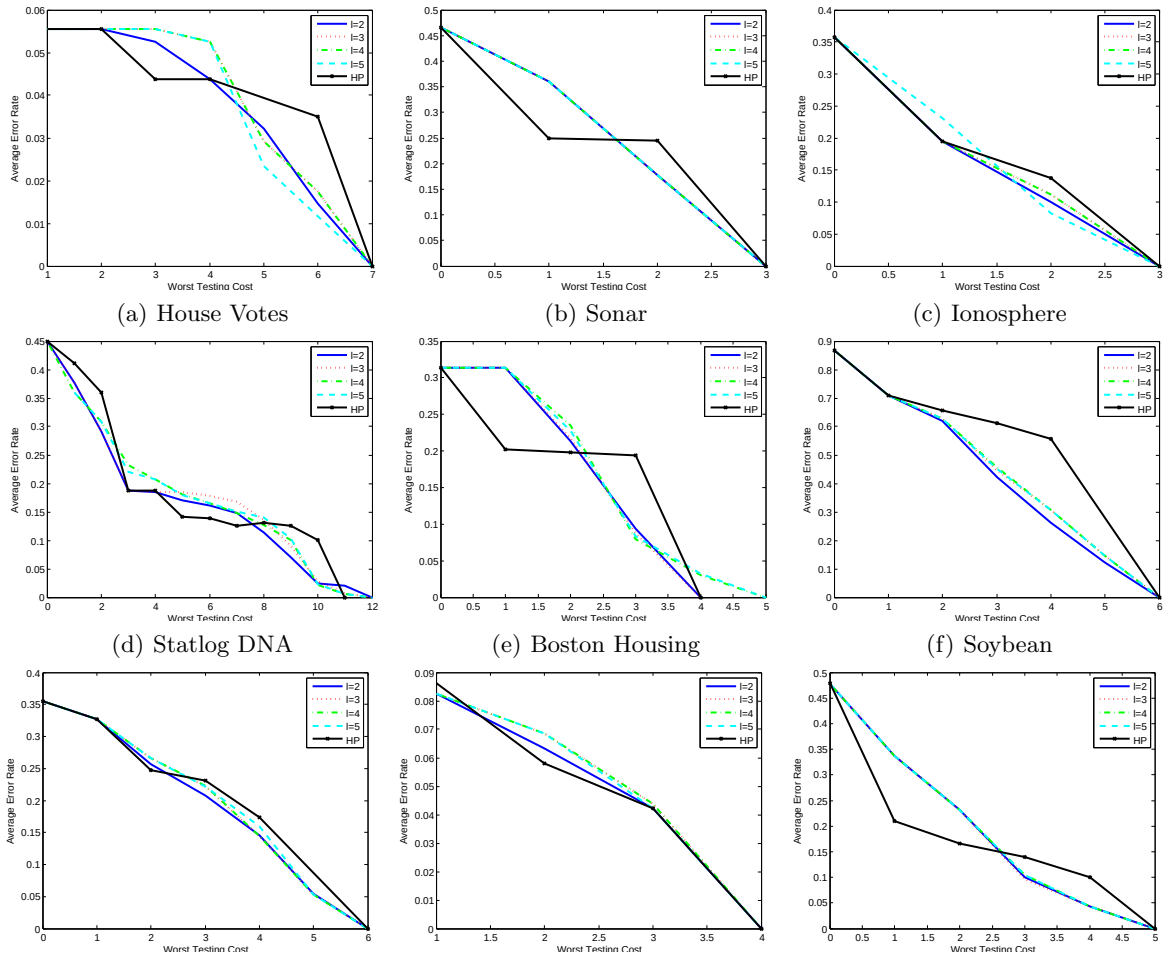
Corollary 0.2. *Powers function*

$$F(G) = (\sum_{i=1}^{k} n_G^i)^l - \sum_{i=1}^{k} (n_G^i)^l \tag{3}$$

is admissible *for* $l = 2, 3, \dots$.

We compare the threshold-Pairs with various α values against the Powers function to study the effect of them on the tree building subroutine GreedyTree. We compare performance using 9 data sets from the UCI Repository in Figure 1. We assume that all features have a uniform cost. For each data set, we replace non-unique objects with a single instance using the most common label for the objects, allowing every data set to be complete (perfectly classified by the decision trees). Additionally, continuous features are transformed to discrete features by quantizing to 10 uniformly spaced levels. For trees with a smaller cost (and therefore lower depth), the threshold-Pairs impurity function outperforms the Powers impurity function with early stopping (higher α leads to earlier stopping), whereas for larger cost (and greater depth), the Powers impurity function outperforms threshold-Pairs. If α is set to 0, the difference between threshold-Pairs and Powers function is small.

Details of Data Sets The house votes data set is composed of the voting records for 435 members of the U.S. House of Representatives (342 unique voting records) on 16 measures, with a goal of identifying the party of each member. The sonar data set contains 208 sonar signatures, each composed of energy levels (quantized to 10 levels) in 60 different frequency bands, with a goal of identifying The ionosphere data set has 351 (350 unique) radar returns, each composed of 34 responses (quantized to 10 levels), with a goal of identifying if an event represents a free electron in the ionosphere. The Statlog DNA data set is composed of 3186 (3001 unique) DNA sequences with 180 features, with a goal of predicting whether the sequence represents a boundary of DNA to be spliced in or out. The Boston housing data set contains 13 attributes (quantized to 10 levels) pertaining to 506 (469 unique) different neighborhoods around Boston, with a goal of predicting which quartile the median income of the neighborhood the neighborhood falls. The soybean data set is composed of 307 examples (303 unique) composed of 34 categorical features, with a goal of predicting from among 19 diseases which is afflicting the soy bean plant. The pima data set is composed of 8 features (with continuous features quantized to 10 levels) corresponding to medical information and tests for 768 patients (753 unique feature patterns), with a goal of diagnosing diabetes. The Wisconsin breast cancer data set contains 30 features corresponding to properties of a cell nucleus for 569 samples, with a goal of identifying if the cell is malignant or benign. The mammography data set contains 6 features from mammography scans (with age quantized into 10 bins) for 830 patients, with a goal of classifying the lesions as malignant or benign.



(a) House Votes (b) Sonar (c) Ionosphere

(d) Statlog DNA (e) Boston Housing (f) Soybean

(g) Pima (h) WBCD (i) Mammography

Figure 1. Comparison of classification error vs. max-cost for the Powers impurity function in (3) for $l = 2, 3, 4, 5$ and the threshold-Pairs impurity function. Note that for both House Votes and WBCD, the depth 0 tree is not included as the error decreases dramatically using a single test. In many cases, the threshold-Pairs impurity function outperforms the Powers impurity functions for trees with smaller max-costs, whereas the Powers impurity function outperforms the threshold-Pairs function for larger max-costs.