Cost-complexity pruning of random forests

ISMM 2017, 13th International Symposium on Mathematical Morphology, Fontainebleau, France, May 15 - 17, 2017

Why perform pruning?

- Out-of-bag samples are un-used samples from the Bootstrap Aggregation procedure in random forests.
- We study the effect of using the out-of-bag samples to improve the generalization error first of the decision trees, and second the random forest by post-pruning.

Decision Tree Pruning

- Cost-Complexity function: \( g(t) = R(T) - R(T_{t_{\text{Leaves}(T)}}) \) \(|\text{Leaves}(T)| - 1 \)
- \( R(T) = \sum_{t \in \text{Leaves}(T)} r(t) \cdot p(t) = \sum_{t \in \text{Leaves}(T)} R(t) \)
- \( R(T) \) is the training error, \( \text{Leaves}(T) \) is the number of leaves of tree \( T \)
- \( r(t) = 1 - \max_k p(C_k) \) is the misclassification rate and \( p(t) = n_t/N \) is the number of samples in node \( n_t \) to total training samples \( N \).

Random Forests

- Decision tree ensembles: Random Forests (RF), Extremely Randomize Trees (ET), Bagged trees (BT)
- Randomization Methods: Bootstrap Aggregation, Random Feature selection, Random Threshold section

Out-Of-Bag Cost Complexity Pruning

- Independent tree pruning:
  \( T_j^* = \arg\min_{\alpha \in A_j} \mathbb{E} \| Y_{\text{OOB}} - T_j^{(\alpha)}(X_{\text{OOB}}) \|^2 \)
- Global threshold pruning:
  \( \{ T_j^* \}_{j=1}^M = \arg\min_{\alpha \in A_j} \mathbb{E} \| Y_{\text{train}} - \frac{1}{M} \sum_{j=1}^M T_j^{(\alpha)}(X_{\text{OOB}}) \|^2 \)

Results and Analysis

- Plots of \( A_j \forall j \) for RFs, ETs, BTs:
  - Figure: RFs and ETs provide a larger subset of CC-parameter values \( A_j \) and thus subtrees \( T_j \) for the cross-validation step.

Performance on datasets from UCI repository:

- Reduction in forest size for marginal loss in classification accuracy.
- Out-of-Bag samples provide cross-validation mechanism to prune forests.

Future work

- Understand non-monotonicity (spikes) of random forest training error.
- Does post-pruning preserve consistency of forests?
- How to define a global cost-complexity parameter for random forests?