Hoeffding Option Trees

Data streams project

Daniela Pistol
Xiaoxuan Hei

Introduction

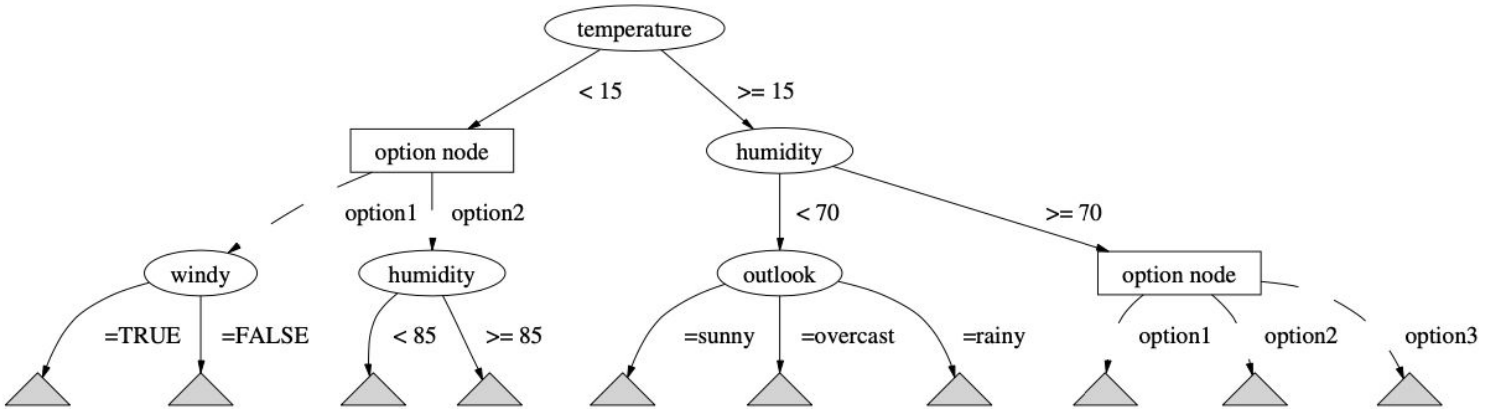
- **Data streams**: a single scan of the data as quickly as possible
- **Hoeffding trees -** incremental decision tree learner for data streams
  - hoeffding bound

$$\epsilon = \sqrt{\frac{R^2 ln(1/\delta)}{2n_l}}$$

- **Hoeffding option trees -** middle ground between single trees and ensembles

Hoeffding Option Trees

- **Simple node -** just one test: rainy or sunny?
- **Option node -** multiple tests → multiple paths/ trees



Managing tree growth

- **Counter**: to limit the number of options per option node
  - after previous experiments and results we set this variable to 5

Restricting additional splits

- only consider options for attributes that have not already been split at that node to prevent redundant subtrees

Algorithm

- **For each sample and for each option node**:
  - Compute the sufficient statistics
  - if a node has no children, add a node which splits on the highest information gain scored attribute
    - add leaves with statistics for each branch
  - if a node has already children, add a node which splits on the attribute with the highest information gain and which has not been used until now
    - add leaves with statistics for each branch

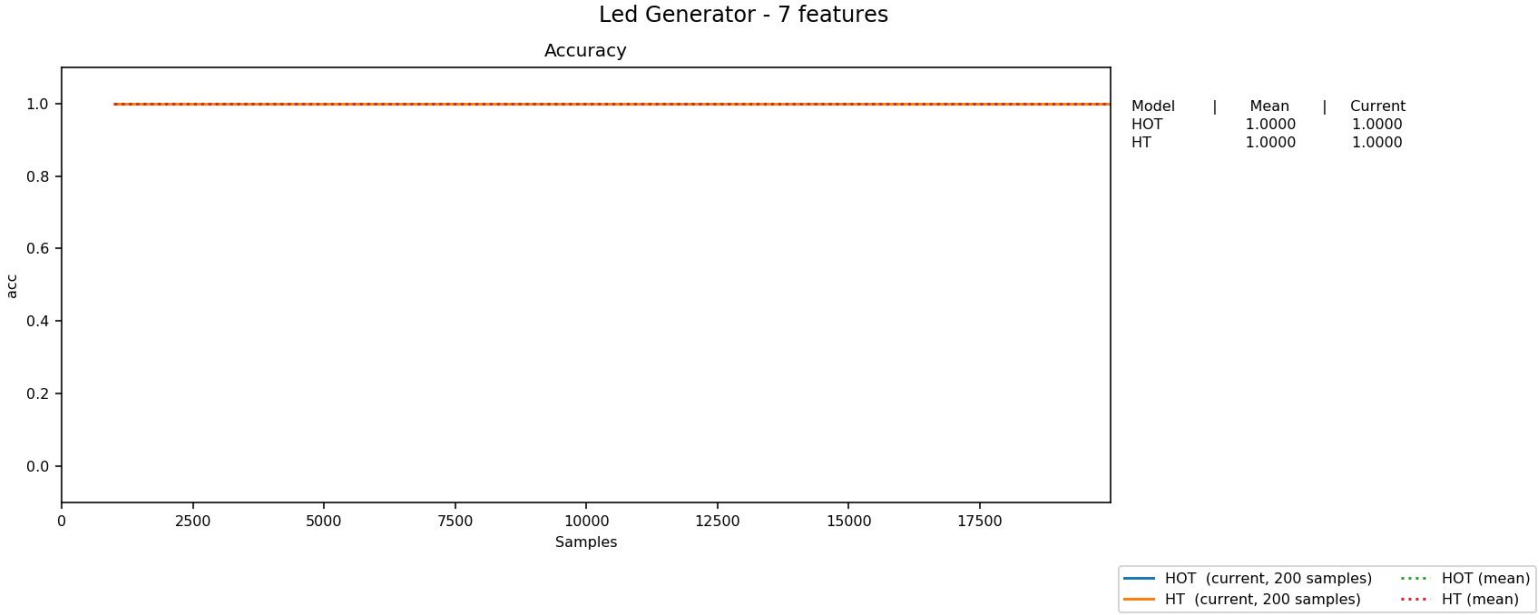
- **Attention! we add the nodes and the branches only if the Hoeffding bound is respected!**

$$\epsilon = \sqrt{\frac{R^2 ln(1/\delta)}{2n_l}}$$

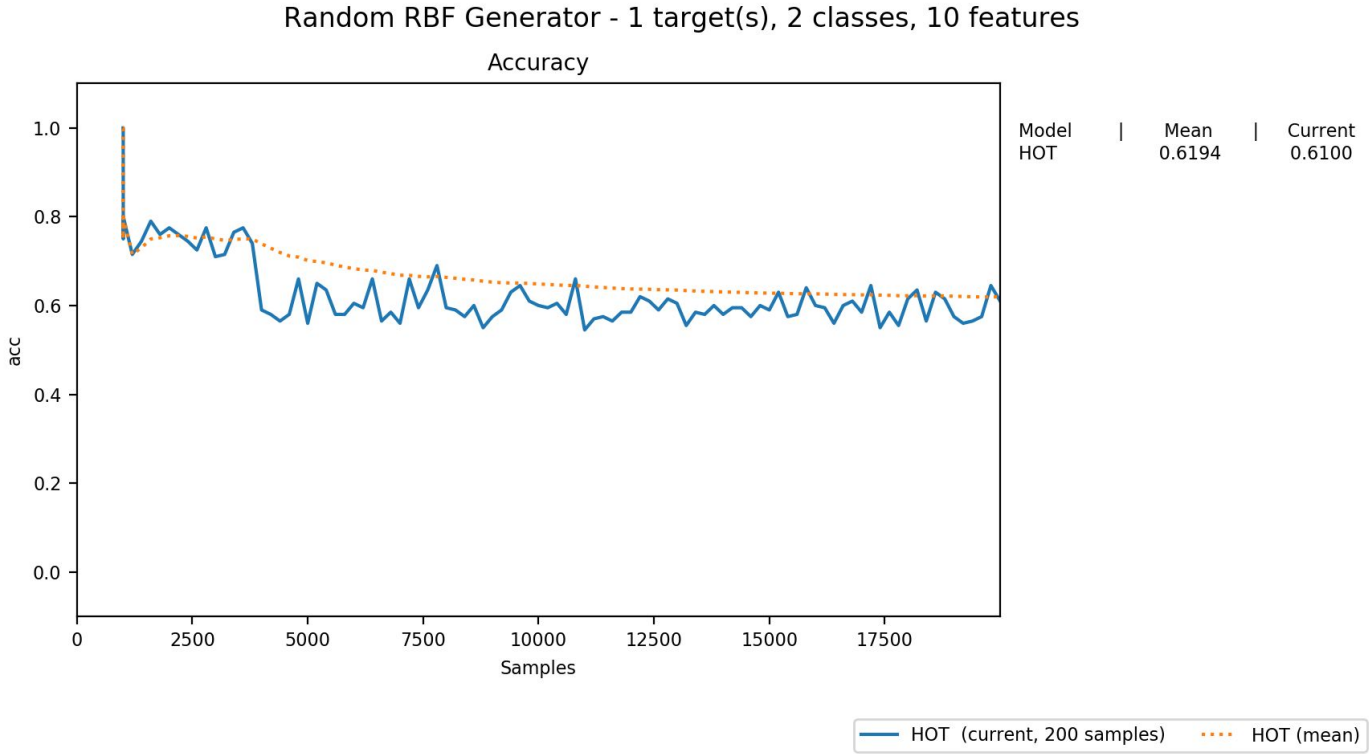
  - R is the range of information gain (highest-lowest)
  - delta is the confidence for additional splits
  - nl is the number of exemples seen at the node l until now

This Hoeffding bound should be lower than the information gain difference between the highest scored attribute and the second one for a node with no children, and lower than the difference between the child and an unseen attribute with the highest information gain.

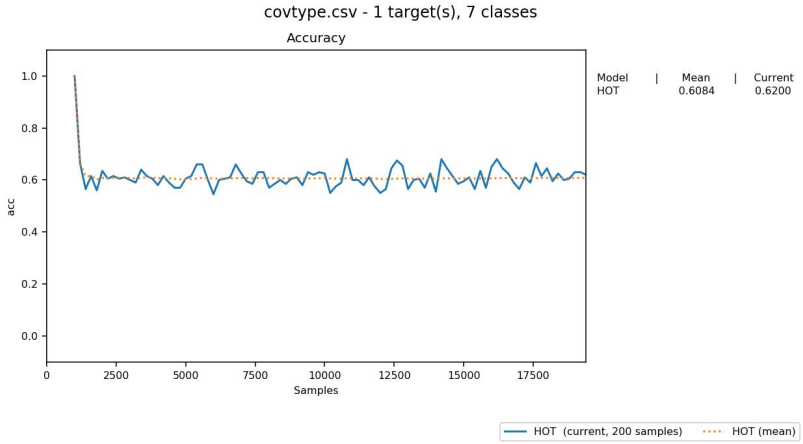
Experiment Results

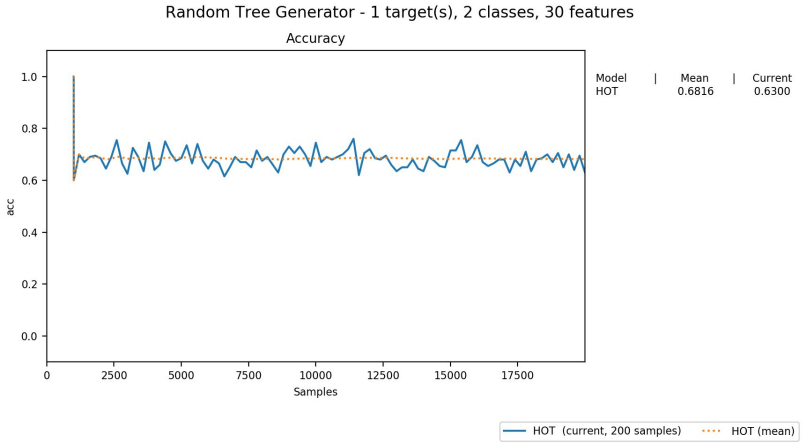


Experiment Results

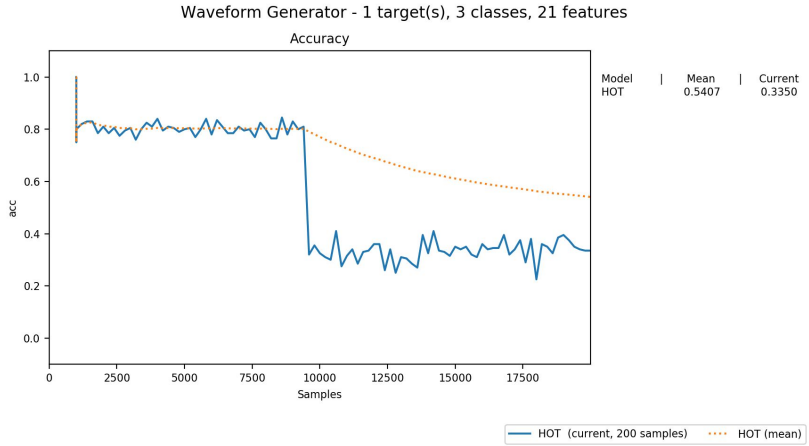


Experiment Results

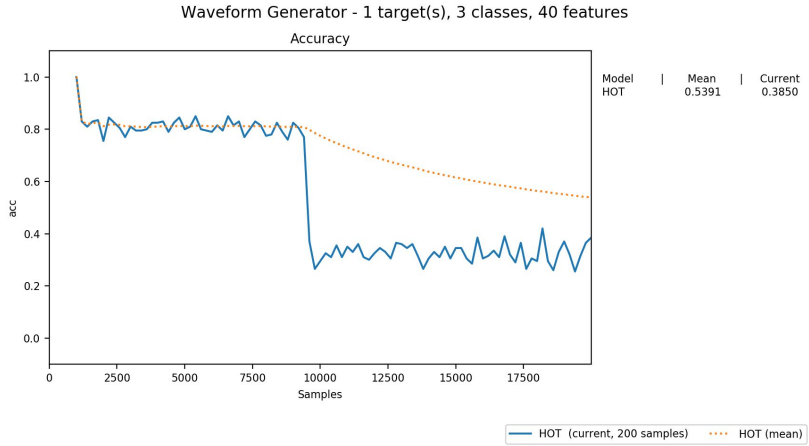




Experiment Results



Wave 21 (Wave without noise)



Wave 40 (Wave with noise)

Experiment Results

	HT		HOT	
Dataset	acc	time	acc	time
LED	100.00	19.26	100.00	16.24
Cover Type	60.78	107.90	60.84	112.36
Random Tree	68.57	16.36	68.16	7.42
RRBF	80.36	11.21	61.94	6.73
Wave 21	78.70	19,47	54.07	13.20
Wave 24	80.02	29.86	53.91	17.62

Experiment Results