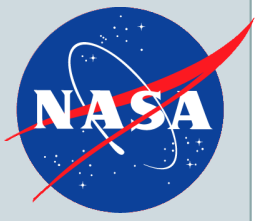
Machine Learning Methods for Astronomy

Kiri Wagstaff, David Thompson, Umaa Rebbapragada
Jet Propulsion Laboratory, California Institute of Technology

Cost-sensitive Learning
Cost: Computation? Time? Features?

Collaborative Analysis
Arrays, sensor networks

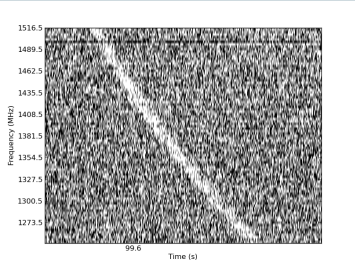
Anomaly Detection



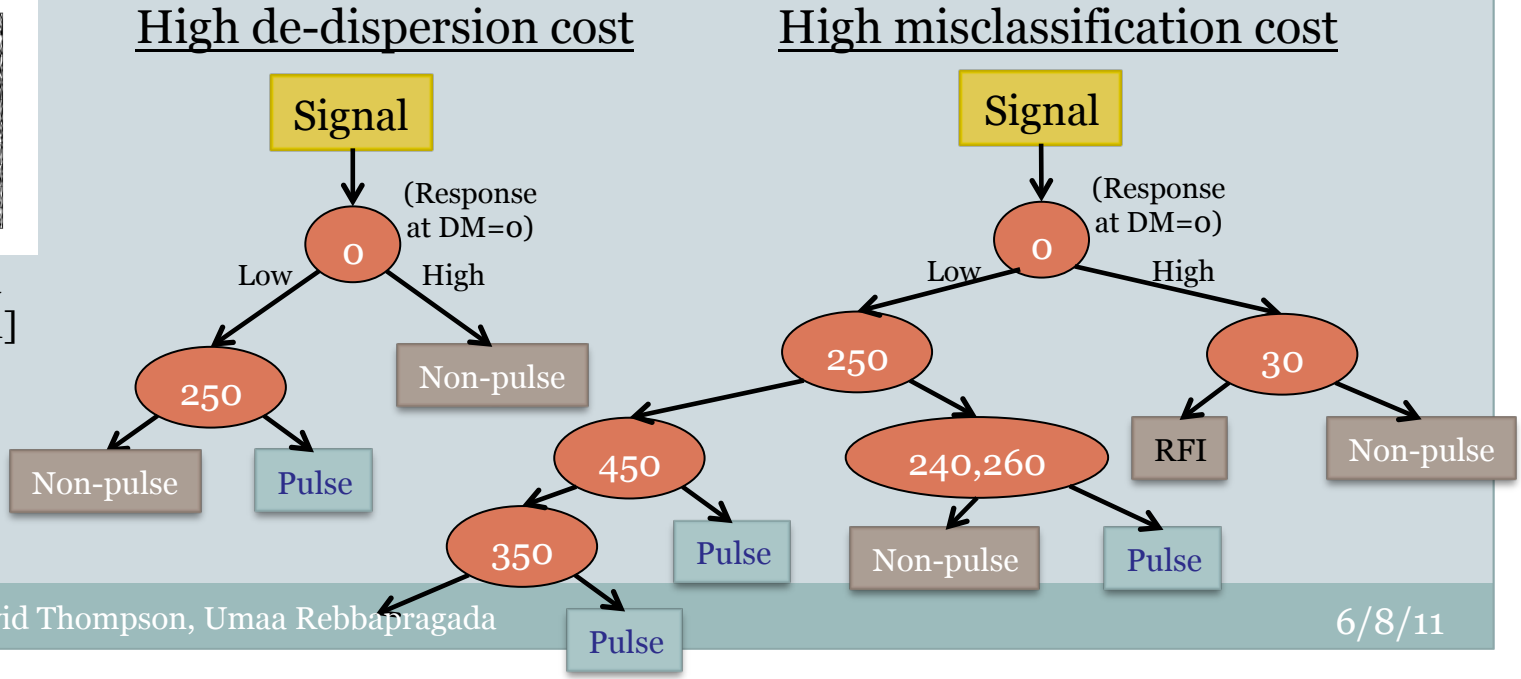
This work was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, © 2011. Government sponsorship acknowledged.

Cost-Sensitive Learning (1)

- Selective computation based on cost
- Cost-sensitive decision tree (CSDT) [Ling et al., 04]
  - Instead of maximizing information gain, build tree to minimize cost of errors + cost of feature acquisition
  - Speculatively: Decision tree nodes = computation, not just lookup

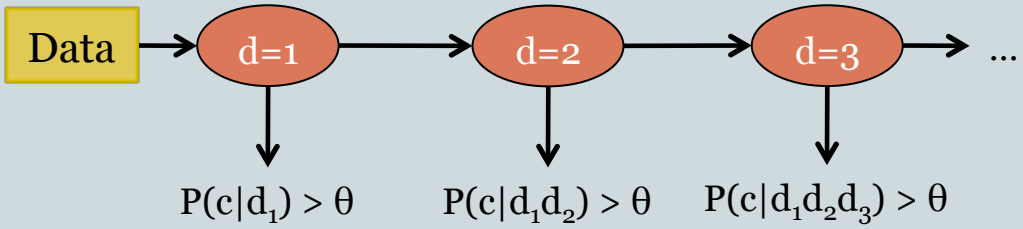


Parkes radio data [Edwards et al., 01]



Cost-Sensitive Learning (2)

- Cost-sensitive Feature Acquisition (CFA); cascade ensemble [desJardins et al., 2010]
  - Build a classifier using "free" features
  - For poorly classified items, acquire another feature and train a new classifier
  - Continue until features exhausted or all items classified well



  - Minimizes acquisition cost while maintaining desired posterior prob.
  - Also: reliable (abstaining) classifiers [Vanderlooy et al., 2009]

Collaborative Analysis (1)

- Ensemble: multi-station transient detection (VLBA)
  - Leverages differences in local RFI environments
  - Assumption: real transients will be detected by more than one station
  - Installed for commensal detection at the VLBA

1. Status quo (incoherent sum)

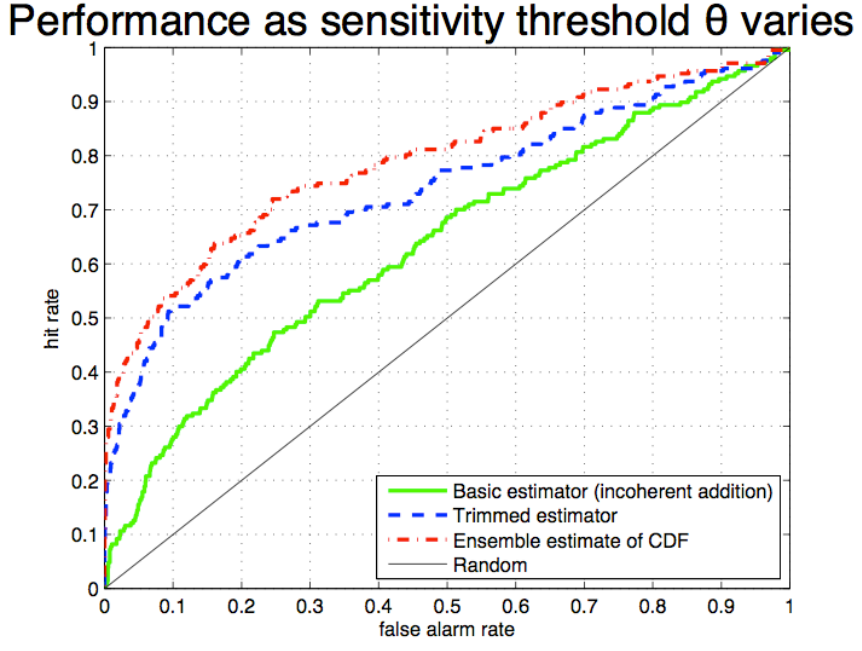
$$\max_{DM} \frac{1}{A} \sum_{a=1}^{A} S_a(DM)$$

2. Trimmed robust estimator

$$\max_{DM} \frac{1}{A} \sum_{a=1}^{A-1} S_a(DM)$$

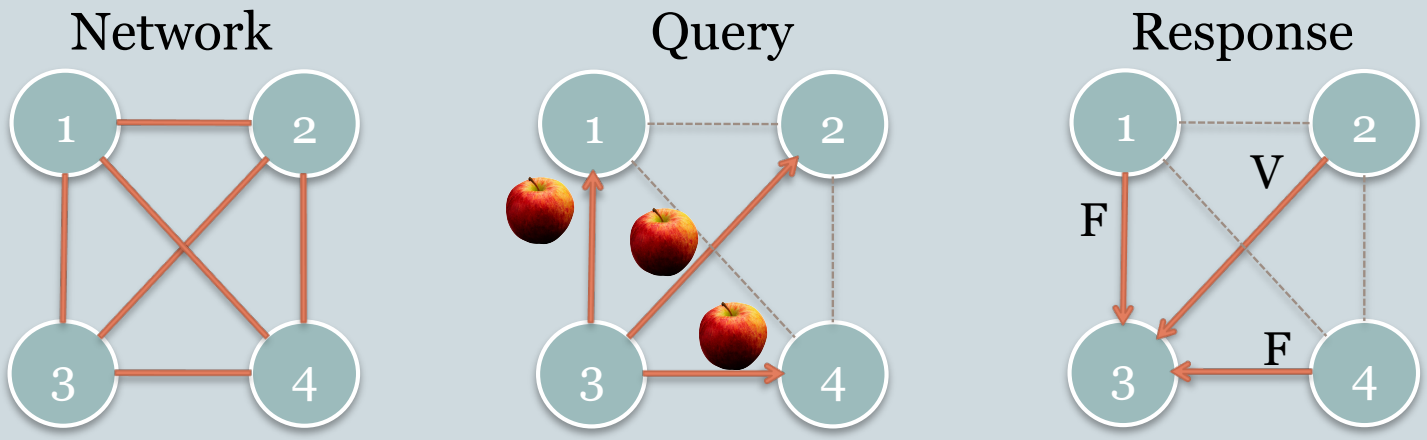
3. Ensemble CDF estimation

$$\max_{DM} \frac{1}{A} \sum_{a=1}^{A} P(X \leq S_a(DM))$$



Collaborative Analysis (2)

- Independent: collaborative classification and clustering
  - Learners bootstrap each other to higher performance (like co-training, [Blum & Mitchell, 1998])
  - Each learner queries neighbors for new data labels (or constraints), shares the result, then retrains [Rebbapragada & Wagstaff, 2011]



  - Enables autonomous learning with minimal human effort

Anomaly Detection

- SSEND: Semi-supervised Eigenbasis Novelty Detection [Thompson et al., submitted]
  - Project data into lower dimensional space (basis) and detect anomalies by their reconstruction error
  - Semi-supervised: include known uninteresting examples (e.g., RFI)