Needle in a Haystack: Label-Efficient Evaluation under Extreme Class Imbalance

Neil Marchant and Ben Rubinstein

School of Computing and Information Systems, University of Melbourne, Australia

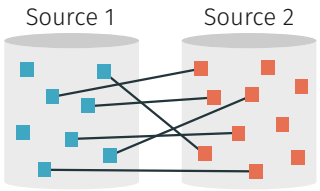
1. Evaluation: who watches the watchmen?

- Classifier evaluation is unreliable when statistical bias and noise are uncontrolled
- Difficult to overcome due to costly ground truth, data imbalance and biased sampling

2. Passive sampling fails under data imbalance

When the performance measure is sensitive to rare instances, a large passive sample is needed to drive down statistical error

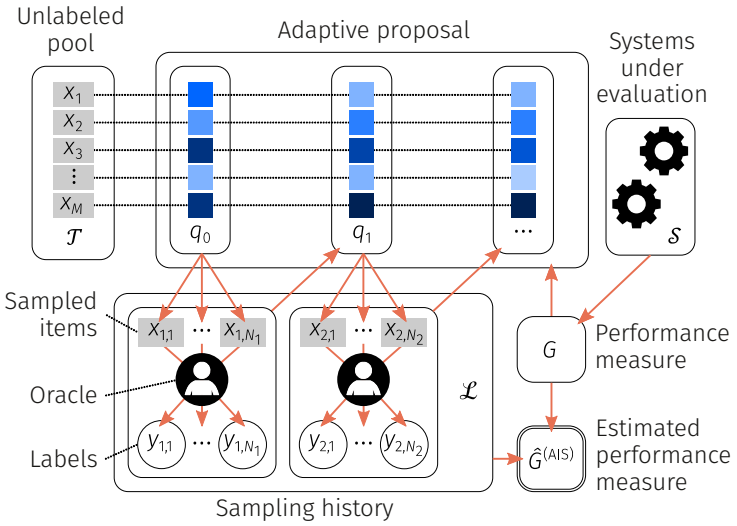
Example: in record linkage finding a match is like finding a needle in a haystack - e.g. 1 match per 1 million non-matches



Other affected domains: rare diseases, risk prediction, extreme classification

3. Our AIS-based evaluation framework

Objective: accurate and precise performance estimates using minimal labeled examples



Given: systems to evaluate; target performance measure; unlabeled pool of examples; labeling oracle (e.g. human expert)

Return: sampling history; estimated performance measure; approx. confidence region (optional)

Design: application of *adaptive importance sampling*—labels collected in batches; items selected to label via an adaptive proposal.

4. Generalized performance measures

Supported performance measures are vector-valued risk functionals R mapped through g

$$G = g(R) \text{ with } R = \mathbb{E}[\ell(X, Y)]$$

Encompasses: classification and regression measures; a vector of measures; PR/ROC curves

5. Theoretical guarantees

Peformance estimates satisfy:

- strong consistency
- a central limit theorem (CLT)

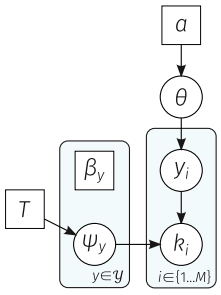
Use the CLT to derive the asymptotically-optimal variance-minimizing proposal

6. Adapting the proposal

Given: online model for the oracle response; sampling history

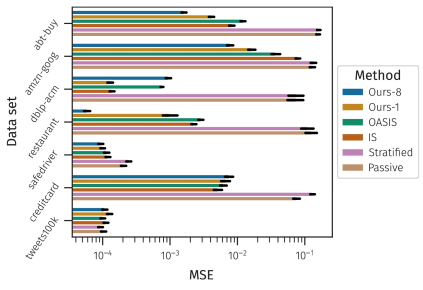
Return: approximation to the asymptotically-optimal proposal

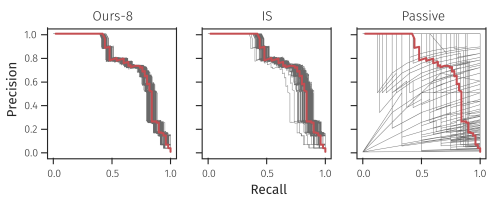
We instantitate with a Dirichlet-tree model that's asymptotically-optimal for a deterministic oracle



7. Experimental results

Best or equal-best MSE on 6 of 7 datasets when estimating F1 score





Also works for vector-valued measures like precision-recall curves

Code: github.com/ngmarchant/ActiveEval

Full tech report: arXiv 2006.06963

N. Marchant acknowledges the support of an Australian Government Research Training Program Scholarship. B. Rubinstein acknowledges the support of Australian Research Council grant DP150103710.