Active Learning for Biomedical Citation Screening

Byron Wallace¹,², Kevin Small¹, Carla Brodley¹ & Thomas Trikalinos²

¹ Department of Computer Science, Tufts University
² Center for Clinical Evidence Synthesis, Tufts Medical Center
Systematic Reviews

**Systematic review**: an exhaustive assessment of all the published medical evidence regarding a precise clinical question.

e.g., “Is aspirin better than leeches in inducing more than 50% relief in patients with tension headaches?”

**Must find all relevant studies.**
Aim: use active learning + domain knowledge to semi-automate the screening step
Citation Screening

Doctors read all these. They’d rather be doing something else.

This is a natural fit for active learning
Caveat
or why ‘off-the-shelf’ AL doesn’t work

Imbalanced data; ‘relevant’ class is very small (≈5%), but sensitivity to this class is paramount
Why Might Uncertainty Sampling Fail?

- Hasty generalization: uncertainty sampling may miss 'clusters'
  - Pre-clustering doesn’t help
    - unreliable in high-dimensions
    - small clusters of interest

Prior knowledge may be a way around this problem
Guiding AL With Domain Knowledge

- Experts bring lots of knowledge to the table
- **Labeled terms**: terms or $n$-grams whose presence is indicative of class membership. In our toy example:

  - $+$ tension headache, leeches, aspirin
  - $-$ migraine headache, mice

  “Is aspirin better than leeches in inducing more than 50% relief in patients with tension headaches?”
Co-Testing Framework
(Muslea et al., 2000)

If model 1 disagrees with model 2 about $\mathbf{x}$, then $\mathbf{x}$ is a good point to label
Labeled Terms + Co-Testing

**Model 1** Standard BOW (linear kernel) SVM

**Model 2** $|\log(\#\text{pos terms} / \#\text{neg terms})|$

Query strategy:

1. Find all documents about which the models disagree
2. Select for labeling $\arg\max_x F_2(x)$
COPD: Genetic Associations with COPD
Micronutrients: Effects on Health
Take Home

Prior knowledge (here in the form of labeled terms) can be used to guide active learning
Other Issues in Deployed AL
(See Our Paper)

• What if the expert can provide ranked labeled features?
  – e.g., features x and y are both indicative of class c, and x is more so than y

• How should AL approaches be evaluated in real-world projects?
  – Costs are asymmetric; but how to quantify this?
  – Finite pool scenarios: we don’t always care about the induced classifiers’ accuracy

• Concept drift is a reality. How to deal with this?
Software & Data

• Our software is currently in use at Tufts Medical Center, for systematic reviews covering the following topics:
  – Diagnostic test assessment
  – Crohn’s disease
  – Sleep apnea
  – Blood pressure

• Code & data available at:
  http://github.com/bwallace/citation-screening