Combining Unsupervised Learning with Active Learning for the Labeling of Large Sample Corpora

Ted Markowitz
Pace University CSIS DPS & IBM T. J. Watson Research Ctr.
tjm@cognosys.net
Agenda

- Quick History
- Research Motivation
- Active Learning
- Selective Sampling
- *Approximate Uncertainty Sampling*
- Current Status
- Lessons Learned
Quick History

- Member of DPS 2006 class
- In 2001 started small IT consulting firm
- Prior to 2001, was senior technologist at:
  - American Express: Dir. R&D
  - Xerox: Chief Knowledge Engineer
  - Netscape / AOL: Principal Consultant
  - RetailDNA (Walker Digital): Chief Architect
• Anti-spam research suggested by Dr. Chuck Tappert’s work with Ian Stuart, MS student
• Decided to approach IBM Research’s SpamGuru anti-spam group to work jointly
• Started P/T onsite at IBM in 11/05—after many months of Pace-IBM contract negotiations
• Dr. Richard Segal of IBM generously agreed to act as de facto DPS advisor
Research Motivation

• **Assumption:** Ongoing training and evaluation of anti-spam tools require fresh, large corpora (databases) of labeled (spam vs. good) email.

• **Problem:** How do we accurately label large numbers of examples—many thousands to millions—without manually examining each and every one?
Email Corpora: Issues

• Existing email corpora are typically:
  – Rather small (few thousand messages)
  – Very narrowly focused in type and content
  – Growing more and more stale over time

• Both good mail & spam continually evolve

• Email environments can look very different

• While it’s easy to collect spam, there are never enough examples of good mail
Email Corpora: Bottom Line

- Building large, current, and diverse training and testing corpora is hard and expensive
- *Result*: Just a few—relatively small and continuously aging—email test corpora are used again and again while the target environments are constantly evolving
One Potential Approach

- Machine Learning (ML) techniques can be used to build classifiers which learn to label.
- ML research has shown it is possible to learn from a small # of labeled examples and lots of unlabeled examples.
- Advanced sampling techniques can be used to minimize manual labeling intervention.
“Active Learning” (AL)

• Unlike passive learning, an Active Learner can direct attention to particular areas of the domain it wants to request information about

• Learner (student) has access to a teacher who can answer questions about all the examples

• AL questions are often phrased in the form of membership queries, e.g., “To what class does this example belong?”
“Active Learning”

- **Challenge**: Maximizing future prediction accuracy while minimizing the total # of queries required to achieve target error rates.
- One major research question in AL is how to select the “best” set of sample queries during each iteration of the learning cycle.
Selective Sampling

- **Uncertainty Sampling (US)** is one technique for selecting the most informative examples.
- US is predicated on the idea that the learner will learn better by asking for those examples about which it is most uncertain.
US Algorithm†

1. Create an initial classifier
2. While the teacher is willing to label examples...
   i. Apply the current classifier to each unlabeled example
   ii. Find the $b$ examples for which the labeler is least certain of class membership
   iii. Have the teacher label the subsample of $b$ examples
   iv. Train a new classifier on all labeled examples

† “A Sequential Algorithm for Training Text Classifiers”, D. D. Lewis & W. A. Gale, ACM SIGIR ‘94
US Issues

- Minimizing total uncertainty over all examples is computationally expensive: $O(n)$
- Can you reduce the number of questions asked in each cycle and still learn accurately?
- Is always picking just the most uncertain examples the best learning strategy?
- What other knowledge can be brought to bear in selecting the best questions?
Research Approach

- Construct competing AL/US-based labelers
- Evaluate them:
  - Compare accuracy (% correct vs. false positives & false negatives)
  - Compare sample sizes required
  - Minimize # of end-user questions required
  - Maximize performance
- Select best labelers and refine them
Research Environment

- Built Java labeler testbench for comparing labeler variations; based on IBM SpamGuru codebase
- Developed and tested several alternative Uncertainty Sampling-based labelers
- Employed pre-labeled TREC 2005 Enron 92K msg corpus to simulate AL teacher
• From TREC corpus created five different 73K msg subsamples to use for cross-validation
• Also used sample msgs from select IBM users
• Built Java GUI front-end (CSI) to support human teacher interaction with labelers
Labeler Testbench

Teacher / Oracle

"Blackbox" Labeler

Labels (ham/spam)

Data (msgs)

Other Classifier(s)

Optional
Research Status

- Focused on development of *Approximate Uncertainty Sampling* (AUS) labeler
  - Compromise between speed of learning, # of questions asked & processing resources required
  - Computational complexity of AUS is < original Uncertainty Sampling algorithm
- Submitted AUS paper to CEAS 2006 anti-spam conference with initial AUS results: **PAPER ACCEPTED!**
FUNCTION SelectQuery AUS(C, U, M, R) returns Q, R

// Input
// C: Classifier trained on previously labeled data
// U: Unlabeled data
// M: Batch size of queries
// R: Previously recorded uncertainty values

// Output
// Q: Set of examples to label
// R: Updated uncertainty values

Let T = M*log(|U|)
Let S = Select the T examples from U that maximize R()
Foreach e in S do
    Let R(e) = ComputeUncertainty(C, e)
EndFor
Let Q = Select the M examples from U that maximize R()
Return Q
END
Benefits of AUS

• Nearly as effective as Uncertainty Sampling, but with substantially lower computational complexity: $O(M \log(U))$ vs. $O(U)$

• Reduced processing cost allows AUS to be applied to labeling larger datasets

• AUS also makes it possible to update the learned model more frequently
Current Work

- Reexamine why query set generation using unsupervised clustering (AUS3 & AUS4) didn’t produce better results
- Work on better clustering versions of AUS to try to beat AUS2 performance
- Continue literature review on Active Learning, Unsupervised Learning, Selective Sampling, and other related ML techniques
A Few Lessons Learned

• It’s really terrific to work with smart people!
• *It can also be tough to work with smart people!*
• Nailing down a topic can be truly arduous, even when you’re focused on a good problem area, operating in a great environment & are working really hard at it
A Few Lessons Learned

• It’s easy to get frustrated: talk through any issues candidly and frequently
• It’s easy to get overwhelmed: focus on shorter-term, achievable goals with dates
• “Real life” will intrude: remember that it will and deal with it, but keep making progress—any progress—no matter how small it seems!