Accelerating Active Learning with Transfer Learning
ICDM 2013

Dave Kale and Yan Liu

University of Southern California
Computer Science Department

December 10, 2013
Active learning challenge: *cold starts* (start with 0 labels)

Test set performance improves very slowly vs. # points, queries.
Fight *cold starts* by starting with labeled data

*Common applied AL heuristic*: seed base learner with labeled data.

What if data comes from different task? Use transfer learning (TL).

Desired properties in an active+transfer learning framework:

- Improve test set performance with few or no target label queries.
- Converge to same or better performance vs. plain AL.
- Theoretically sound.* Build on body of theoretical AL, TL research.
- Easy to implement and understand, flexible, fast.

* "Anything which is inconsistent isn’t a complete solution.”
- email from prominent AL researcher
Consistent online active learning

**Goal:** learn target classification task $T$ in *online* setting
**Input:** stream of target data, target *oracle*

**Algorithm:** Online IWAL CAL (original paper: [1])

Let $H_1 := H$

For $t = 1, \ldots$

Compute $G_t$: disagreement about $h(x_t)$ among $h \in H_t$

Flip coin with $P(\text{heads}) = p_t \approx O(\min\{1, 1/G_t\})$

If heads, then query label $y_t$, set $w_t := 1/p_t$
else set $y_t := 0$, $w_t := 0$ (i.e., ignore $x_t$)

Let $H_{t+1} = \{h : h \text{ consistent with } (x_i, y_i, w_i) \text{ for } i = 1, \ldots, t\}$

- $G_t := \bar{\epsilon}_{t-1}(\bar{h}'_{t-1}) - \bar{\epsilon}_{t-1}(\bar{h}_{t-1})$ where $\bar{h}'(x_t) \neq \bar{h}(x_t)$
- Importance weighted error: $\bar{\epsilon}_t(h) = \frac{1}{t} \sum_{i=1}^{t} w_i 1(h(x_i) = y_i)$
Combining transfer and active learning

Use IWAL CAL with a convex combination of empirical risks [2]. $\bar{\epsilon}_t$ is importance weighted empirical target error, $\hat{\epsilon}_S$ is source empirical error.

**Definition (Combined weighted empirical risk)**

For hypothesis $h \in \mathcal{H}$, $\alpha \in [0, 1]$, after seeing $t$ target points, let

$$\bar{\epsilon}_{\alpha,t,m}(h) \triangleq \alpha \bar{\epsilon}_t(h) + (1 - \alpha)\hat{\epsilon}_S(h)$$

$$= \frac{\alpha}{t} \sum_{i=1}^{t} w_i 1(h(x_i) \neq y_i) + \frac{1 - \alpha}{m} \sum_{j=1}^{m} 1(h(x_j) \neq y_j)$$

$$= \frac{1}{m + t} \sum_{i=1}^{m+t} v_i 1\{h(x_i) \neq f(x_i)\}$$

$$v_i = \begin{cases} 
(1 - \alpha)(m + t)/m & i \leq m \text{ (source)} \\
\alpha(m + t)/(tp_i) & i > m \text{ (target), labeled} \\
0 & i > m \text{ (target), unlabeled}
\end{cases}$$
Consistent online transfer active learning

Substitute combined weighted error for original importance weighted error.

**Goal:** learn target classification task $T$ in online setting

**Input:** $m$ labeled source data, stream of target data, target oracle

**Algorithm:** Online TIWAL CAL

Let $\mathcal{H}_1 := \{h : h(x) = y \text{ for all source } x \land h \in \mathcal{H}\}$

For $t = 1, \ldots$

Compute $G'_t$: disagreement about $h(x_t)$ among $h \in \mathcal{H}_t$

Flip coin with $P(\text{heads}) = p_t \approx O(\min\{1, 1/G_t\})$

If heads, then query label $y_t$, set $w_t := 1/p_t$

else set $y_t := 0$, $w_t := 0$ (i.e., ignore $x_t$)

Let $\mathcal{H}_{t+1} = \{h : h \text{ consistent with } (x_i, y_i, w_i) \text{ for } i = 1, \ldots, t\}$

- $G'_t := \bar{\epsilon}_{\alpha,t-1,m}(\bar{h}'_{t-1}) - \bar{\epsilon}_{\alpha,t-1,m}(\bar{h}_{t-1})$ where $\bar{h}'(x_t) \neq \bar{h}(x_t)$
How good are the classifiers we learn?

Theorem (Upper bound on target generalization error)

For $\bar{h}_t = \arg\min_{h \in H} \bar{\epsilon}_{\alpha,t,m}(h)$, this holds with probability at least $1 - \delta$:

$$\epsilon_T(\bar{h}_t) \leq \epsilon_T(h_T^*) + \alpha\tilde{O}\left(\frac{C_0 \log t}{t}\right) + (1 - \alpha)\tilde{O}\left(\frac{C_0}{m} + d(S, T) + \epsilon_{ST}^*\right)$$

Decomposes into two error terms, one each active and transfer learning:

- **AL part**: $\tilde{O}\left(\frac{C_0 \log t}{t}\right)$ shrinks as $t$ grows
- **TL part**: constant; depends $m$ and source/target similarity
  - $d(S, T)$: distance between source, target distributions (we use $d_{\mathcal{H} \Delta \mathcal{H}}$ distance [2]; can approximate with *domain separator hypothesis*)
  - $\epsilon_{ST}^* = \min_h \epsilon_S(h) + \epsilon_T(h)$ (assumed to be negligible)
- Trade off using $\alpha$ parameter
  - Small $\alpha$: when tasks similar, $m$ large; reduces number of queries
  - Large $\alpha$: when tasks different or $m$ small; behaves like IWAL CAL
Test set error vs. # label queries: 20 Newsgroups

Test set accuracy vs. # queries for 20 Newsgroups (BvP)

Mean test set accuracy vs. mean # queries for 20 Newsgroups (BvP)

- Fully labeled SL
- IWAL CAL
- Fully labeled TL (HvR)
- TIWAL CAL (HvR)
- Fully labeled TL (AvC)
- TIWAL CAL (AvC)
Test set error vs. # label queries: sentiment [3]

Test set accuracy vs. # queries for sentiment (kitchen)

- Fully labeled SL
- IWAL CAL
- Fully labeled TL (electronics)
- TIWAL CAL (electronics)
- Fully labeled TL (dvd)
- TIWAL CAL (dvd)
Conclusions

Transfer learning can “accelerate” active learning, address cold starts!

Our framework

- **works!** Trivial to extend to pool-based active learning.
- is **simple**. Complicated transfer learning not required.
- is **easy to implement** (once you understand IWAL CAL).
- is **theoretically sound**. Yields insights into problem and applications.

Our work provides a foundation for future work:

- More extensive experimentation. Apply to further problems, data sets.
- Adapt transfer as see more target data [4] [5].
- Experiment with more aggressive active learning algorithms.
- Other problems (e.g., regression), types of queries (e.g., features)
- Try more “extreme” transfer learning (e.g., different features).
This is a great topic to work on!

Human beings use active+transfer learning every day.


“...unlabeled data can suggest learning biases that may improve performance over supervised learning, especially when labeled data are few...A basic observation is that active learning provides the opportunity to validate or refute these biases using label queries, and also to subsequently revise them. Thus, it seems that active learners ought to be able to pursue learning biases much more aggressively than passive learners. A few works on cluster-based sampling and multi-view active learning have appeared, but much remains to be discovered.”
References


The end!

Thanks for listening!

Special thank yous to

- USC Melady Lab, especially Taha Bahadori and Marjan Ghazvininejad
- Sanjoy Dasgupta (UCSD), John Langford (MSR), Byron Wallace (Brown)

Ask me about

- The three years I spent working at Children’s Hospital LA
- *Meaningful Use of Complex Medical Data*: http://mucmd.org
- AAAI 2014 workshop on Artificial Intelligence for health