Active Nearest Neighbors in Changing Environments

Ruth Urner

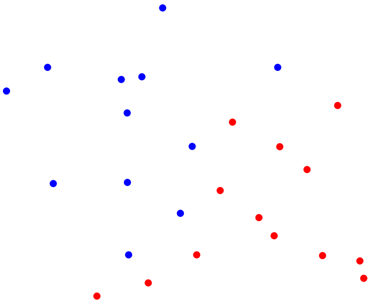
MAX-PLANCK-GESELLSCHAFT

MPI for Intelligent Systems, Tübingen

February 16, 2017

Phenomenon:

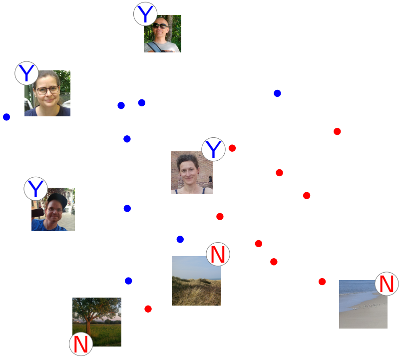
Data generation may change



Active Learning for Domain Adaptation

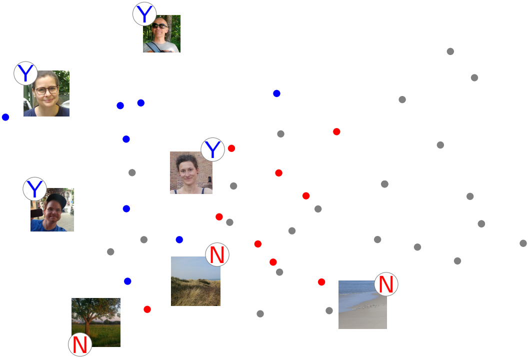
Phenomenon:

Data generation may change



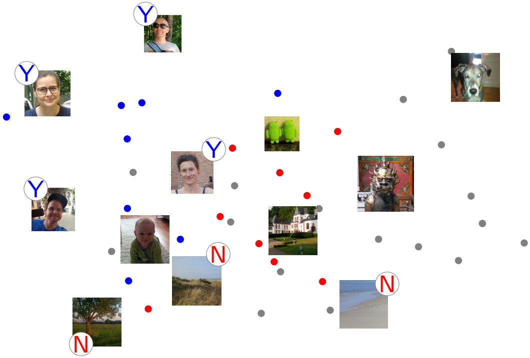
Active Learning for Domain Adaptation

Phenomenon:
Data generation may change



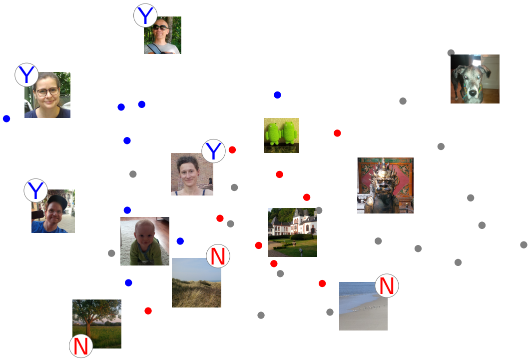
Active Learning for Domain Adaptation

Phenomenon:
Data generation may change



Active Learning for Domain Adaptation

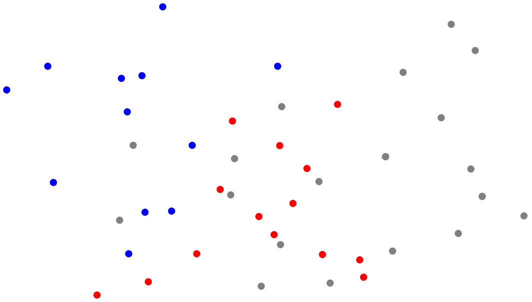
Phenomenon:
Data generation may change



Berlind, U., ICML '15:

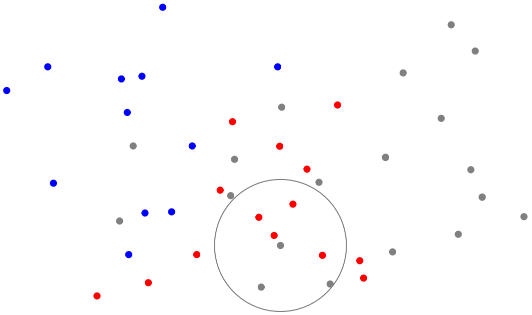
- Developed new learning method ANDA
- **Idea:** use active learning to adapt to distributional shift
- Error bounds on shifted task
- Bounds on number of label queries

Algorithm ANDA:
Nearest Neighbor query rule + Nearest Neighbor prediction



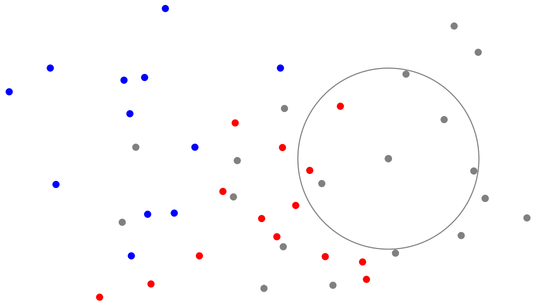
Input: Labeled source data and unlabeled target data

Algorithm ANDA:
Nearest Neighbor query rule + Nearest Neighbor prediction



(k, k')-query rule: don't query!

Algorithm ANDA:
Nearest Neighbor query rule + Nearest Neighbor prediction



(k, k')-query rule: query!

(k, k')-Nearest Neighbor Cover

$T \subseteq \mathcal{X}$, T finite
$k, k' \in \mathbb{N}$ with $k \leq k'$

A set R is a (k, k')-NN-cover for T, if for every $x \in T$, either $x \in R$ or there are k elements from R among the k' nearest neighbors of x in $T \cup R$, that is $|k'(x, T \cup R) \cap R| \geq k$.

ANDA

input: Labeled set S, unlabeled set T, parameters k, k'

- Find $T^l \subseteq T$ s.t. $S \cup T^l$ is a (k, k')-NN-cover of T
- Query the labels of points in T^l

output: $h^k_{S \cup T^l}$, the k-NN classifier on $S \cup T^l$

Lemma

Let T be a finite set of points in a metric space $(\mathcal{X}, \rho)$ and let R be a (k, k')-NN-cover for T. Then, for all $x \in \mathcal{X}$ we have

$$\rho(x, x_k(x, R)) \leq 3\rho(x, x_{k'+1}(x, T))$$

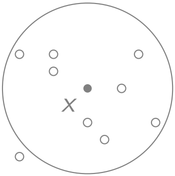
$\Rightarrow$ For every x: the distance to the k nearest labels is at most 3 times the distance to the $k' + 1$ nearest target points!

Proof of Lemma

For every x: the distance to the k nearest labels is at most 3 times the distance to the $k' + 1$ nearest target points.

Proof of Lemma

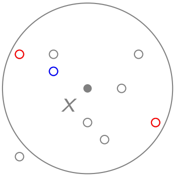
For every x: the distance to the k nearest labels is at most 3 times the distance to the $k' + 1$ nearest target points.



- Let $x \in \mathcal{X}$
- Consider k' nearest neighbors in T

Proof of Lemma

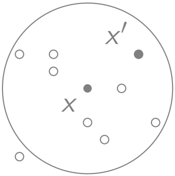
For every x: the distance to the k nearest labels is at most 3 times the distance to the $k' + 1$ nearest target points.



- Let $x \in \mathcal{X}$
- Consider k' nearest neighbors in T
- If they contain k labels $\Rightarrow$ done!

Proof of Lemma

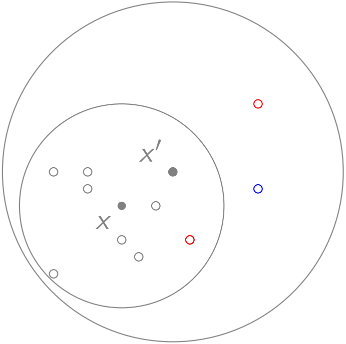
For every x: the distance to the k nearest labels is at most 3 times the distance to the $k' + 1$ nearest target points.



- Let $x \in \mathcal{X}$
- Consider k' nearest neighbors in T
- If they contain k labels $\Rightarrow$ done!
- Else let x' be unlabeled

Proof of Lemma

For every x: the distance to the k nearest labels is at most 3 times the distance to the $k' + 1$ nearest target points.



- Let $x \in \mathcal{X}$
- Consider k' nearest neighbors in T
- If they contain k labels $\Rightarrow$ done!
- Else let x' be unlabeled
- Since x' in T, x' has to be covered!

Error bound

Let $(\mathcal{X}, \rho)$ be a metric space and let P_T be a (target) distribution over $\mathcal{X} \times \{0, 1\}$ with λ-Lipschitz regression function η. Then for all $k' \geq k \geq 10$, all $\epsilon > 0$, and any unlabeled sample size m_T and labeled sequence $S = ((x_1, y_1), \ldots, (x_{m_S}, y_{m_S}))$ with labels y_i generated by η,

$$\mathop{\mathbb{E}}_{T \sim P_T^{m_T}} [\mathcal{L}_T(\mathrm{ANDA}(S, T, k, k'))] \leq \left(1 + \sqrt{\frac{8}{k}}\right) \mathcal{L}_T(h^*) + 9\lambda\epsilon + \frac{2\,\mathrm{N}_\epsilon(\mathcal{X}, \rho)\,k'}{m_T}.$$

Correctness of ANDA does not depend on relatedness assumptions of source and target marginals

However, the number of queries ANDA makes does depend on a local relatedness measure.

However, the number of queries ANDA makes does depend on a local relatedness measure.

Define weight ratio of $B \subseteq \mathcal{X}$:

$$\beta(B) := D_S(B)/D_T(B)$$

Query bound

Let $\delta > 0$, $w > 0$ and $C > 1$. Let m_T be some target sample size with $m_T > k' = (C+1)k$ for some k that satisfies $k \geq 9\,(\mathrm{VC}(\mathcal{B})\ln(2m_T) + \ln(6/\delta))$. Let the source sample size satisfy

$$m_S \geq \frac{36\,\ln(6/\delta) m_T}{C\,w} \ln\left(\frac{9\,m_T}{C\,w}\right)$$

Then, with probability at least $1 - 2\delta$ over samples S of size m_S (*i.i.d.* from P_S) and T of size m_T (*i.i.d.* from D_T), ANDA-S on input S, T, k, k' will not query any points $x \in T$ with $\beta(B_{Ck,T}(x)) > w$.

Query bound provides fall-back guarantee for the lucky case:
If source and target are the same (or very similar/have bounded weight ratio) ANDA will not query at all.

Query consistency

For a fixed target sample size, we show that in the limit of large source samples, ANDA will not make any queries in the support of the source distribution.

Lower bound

- Error bound in terms of Lipschitzness λ and covering numbers $N_{1/\lambda}$
- Query guarantee no queries in source covered area $\mathcal{X}_S \cap \mathcal{X}_T$

Define source coverage of task: $\nu = D_T(\mathcal{X}_S \cap \mathcal{X}_T)$
$\mathcal{C}_\lambda^\nu$: DA tasks with source coverage ν and Lipschitzness λ

Lower bound

Let $(\mathcal{X}, \rho)$ be a metric space, $\nu \in [0,1]$, and $\lambda > 0$. Then for every DA learning algorithm $\mathcal{A}$, every source sample size m_S and target sample size m_T, if $\mathcal{A}$ is restricted to making fewer than

$$q = \frac{\lfloor (1-\nu) Q_{\frac{1}{\lambda}}(\mathcal{X}, \rho) \rfloor}{2}$$

label queries, then there exists a pair of distribution $(P_S, P_T) \in \mathcal{C}_\lambda^\nu$ such that

$$\mathop{\mathbb{E}}_{S \sim P_S^{m_S}, T \sim D_T^{m_T}} [\mathcal{L}_T(\mathcal{A}(S, T))] \geq \frac{1}{4} D_T(\mathcal{X}_T \setminus \mathcal{X}_S)$$

Active learning is beneficial to Domain Adaptation

Corollary
No DA learner with a fixed query budget, in particular no passive DA learner, is consistent on the class $\mathcal{C}^0_\infty$.

But ANDA is :)

Summary

- New method for learning under data shift
- Finite sample bounds on target generalization error

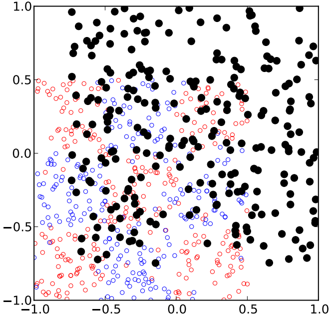
$$\mathop{\mathbb{E}}_{T\sim P_T^{m_T}}[\mathcal{L}_T(\mathrm{ANDA}(S,T,k,k'))] \leq \left(1+\sqrt{\frac{8}{k}}\right)\mathcal{L}_T(h^*) + 9\lambda\epsilon + \frac{2\,\mathrm{N}_\epsilon(\mathcal{X},\rho)\,k'}{m_T}.$$

  (independent of source/target relatedness)

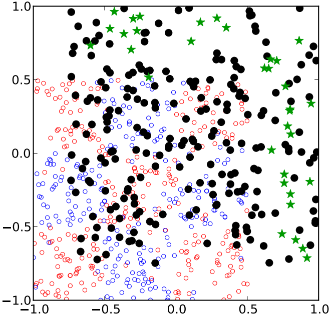
- Adaptability with no prior knowledge of relatedness
- Consistency even when target not supported by the source
- No queries at all when source/target are the same (or similar)

Thank you!

Experiments

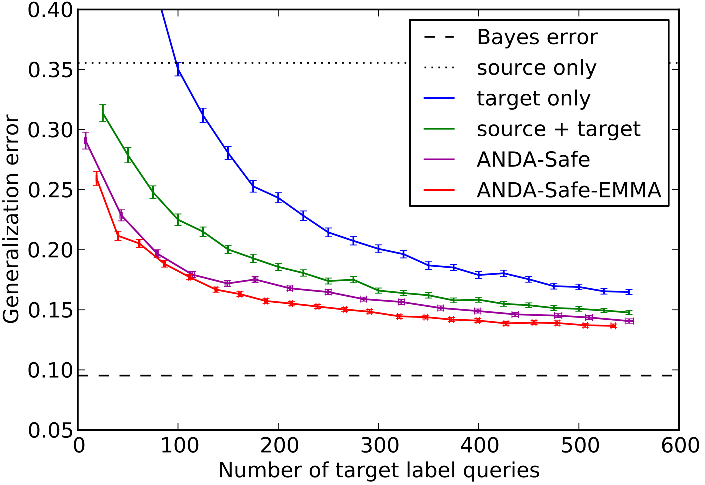


Unlabeled target



Queries made

Experiments



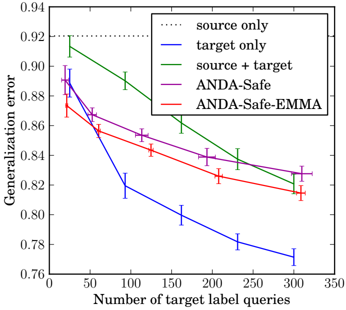


Figure: Imagenet → Caltech256

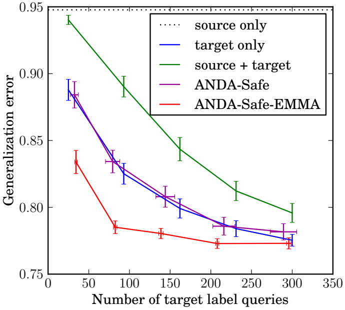


Figure: Bing → Caltech256

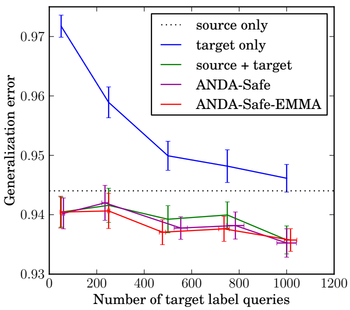


Figure: Caltech256 → Bing

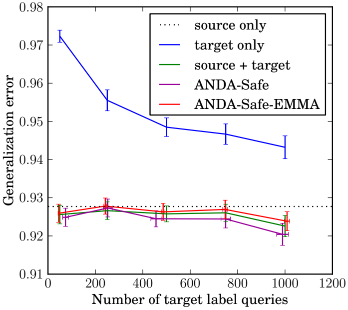


Figure: Imagenet → Bing

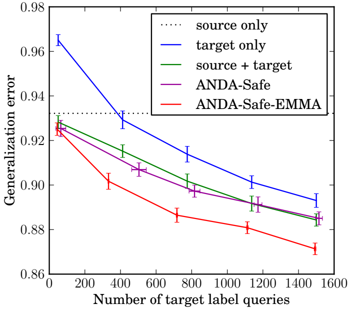


Figure: Bing → Imagenet

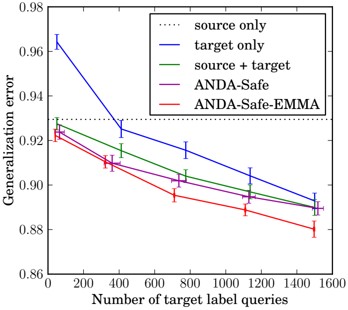


Figure: Caltech256 → Imagenet