Active Nearest Neighbors in Changing Environments

Christopher Berlind CBERLIND@GATECH.EDU
Georgia Institute of Technology, Atlanta, GA, USA

Ruth Urner RUTH.URNER@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

While classic machine learning paradigms assume training and test data are generated from the same process, domain adaptation addresses the more realistic setting in which the learner has large quantities of labeled data from some *source* task but limited or no labeled data from the *target* task it is attempting to learn. In this work, we give the first formal analysis showing that using active learning for domain adaptation yields a way to address the statistical challenges inherent in this setting. We propose a novel nonparametric algorithm, ANDA, that combines an active nearest neighbor querying strategy with nearest neighbor prediction. We provide analyses of its querying behavior and of finite sample convergence rates of the resulting classifier under covariate shift. Our experiments show that ANDA successfully corrects for dataset bias in multiclass image categorization.

1. Introduction

Most machine learning paradigms operate under the assumption that the data generating process remains stable. Training and test data are assumed to be from the same task. However, this is often not an adequate model of reality. For example, image classifiers are often trained on previously collected data before deployment in the real world, where the instances encountered can be systematically different. An e-commerce company may want to predict the success of a product in one country when they only have preference data on that product from consumers in a different country. These and numerous other examples signify the importance of developing learning algorithms that adapt to and perform well in changing environments. This is usually referred to as transfer learning or domain adaptation.

In a common model for domain adaptation, the learner receives large amounts of labeled data from a *source* distribution and unlabeled data from the actual *target* distribution (and possibly a small amount of labeled data from the target task as well). The goal of the learner is to output a good model for the target task. Designing methods for this scenario that are statistically consistent with respect to the target task is important, yet challenging. This difficulty occurs even in the so-called *covariate shift* setting, where the change in the environments is restricted to the marginal over the covariates, while the regression functions (the labeling rules) of the involved distributions are identical.

In this work, we give the first formal analysis showing that using active learning for domain adaptation yields a way to address these challenges. In our model, the learner can make a small number of queries for labels of target examples. Now the goal is to accurately learn a classifier for the target task while making as few label requests as possible. We design and analyze an algorithm showing that being *active adaptive* can yield a consistent learner that uses target labels only where needed.

We propose a simple nonparametric algorithm, ANDA, that combines an active nearest neighbor querying strategy with nearest neighbor prediction. ANDA receives a labeled sample from the source distribution and an unlabeled sample from the target task. It first actively selects a subset of the target data to be labeled based on the amount of source data among the k' nearest neighbors of each target example. Then it outputs a k-nearest neighbor classifier on the combined source and target labeled data.

We prove that ANDA enjoys strong performance guarantees. We first provide a finite sample bound on the expected loss of the resulting classifier in the covariate shift setting. Remarkably, the bound does not depend on source-target relatedness; it only depends on the size of the given unlabeled target sample and properties of the target distribution. This is in stark contrast to most theoretical results for do-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

main adaptation, where additive error terms describing the difference between the source and target frequently appear.

On the other hand, the number of target label queries ANDA makes does depend on the closeness of the involved tasks. ANDA will automatically adjust the number of queries it makes based on local differences between the source and target. We quantify this by giving sample sizes sufficient to guarantee that ANDA makes no queries at all in regions with large enough relative source support. Simply put, ANDA is guaranteed to make enough queries to be consistent but will not make unnecessary ones.

ANDA's intelligent querying behavior and its advantages are further demonstrated by our visualizations and experiments. We visually illustrate ANDA's query strategy and show empirically that ANDA successfully corrects for dataset bias in a challenging image classification task.

1.1. Summary of Main Contributions

The active nearest neighbor algorithm. ANDA operates on a labeled sample from the source distribution and an unlabeled sample from the target task and is parametrized by two integers k and k'. The query rule is to ensure that every target example has at least k labeled examples among its k' nearest neighbors. We describe this formally by defining the concept of a (k, k')-NN-cover, which may be of independent interest for nearest neighbor methods.

Bounding the loss. Theorem 1 provides a finite sample bound on the expected 0-1 loss of the classifier output by ANDA. This bound depends on the size of the unlabeled target sample, the Lipschitz constant of the regression function, and the covering number of the support of the target distribution. It does not depend on the size or the generating process of the labeled source sample. In particular, it does not depend on any relatedness measure between the source and target data generating distributions. We also show that, even dropping the Lipschitz condition, ANDA is still consistent (Corollary 1).

Bounding the number of queries. In Theorem 2 we show that, with high probability, ANDA will not make any queries on points that are sufficiently represented by the source data. This implies in particular that, if the source and target happen to be very similar, ANDA will not make any queries at all. We also prove a "query consistency" result. Together with the error consistency, this implies we get the desired behavior of our active adaptive scheme: its loss converges to the Bayes optimal while queries are made only in regions where the source is uninformative.

Finding a small (k, k')-NN-cover. In general, there are many possible (k, k')-NN-covers to which our theory applies, so finding a *small* cover will use fewer labels for the same error guarantee. We show that finding a minimum-size cover is a special case of the MINIMUM MULTISET MULTICOVER problem (Rajagopalan & Vazirani, 1993). We employ a greedy strategy to find a small cover and argue that it enjoys an $O(\log m)$-approximation guarantee on a combined source/target sample of m points.

Image classification experiments. We demonstrate ANDA's effectiveness in practice by applying it to the problem of dataset bias in image classification. On a collection of 40-class, 1000-dimension datasets (Tommasi & Tuytelaars, 2014), ANDA consistently outperforms baseline nearest neighbor methods, despite the data's high dimensionality and lack of strict adherence to covariate shift. This also shows that ANDA performs well even when our theory's assumptions are not exactly satisfied.

The idea of incorporating active learning (selective querying strategies) into the design of algorithms for domain adaptation has recently received some attention (Chattopadhyay et al., 2013a;b; Saha et al., 2011). However, to the best of our knowledge, there has not been any formal analysis of using active learning to adapt to distribution changes. We believe active learning is a powerful and promising tool for obtaining domain adaptive learners and that this area deserves a sound theoretical foundation. We view our work as a first step in this direction.

1.2. Related Work

There is a rich body of applied studies for transfer or domain adaptation learning (Pan & Yang, 2010), and on selective sampling or active learning (Settles, 2010). We here focus on studies that provide performance guarantees.

For domain adaptation, even under covariate shift, performance guarantees usually involve an extra additive term that measures the difference between source and target tasks (Ben-David et al., 2006; Mansour et al., 2009), or they rely on strong assumptions, such as the target support being a subset of the source support and the density ratio between source and target being bounded from below (Sugiyama et al., 2008; Ben-David & Urner, 2014; Shi & Sha, 2012). Generally, the case where the target is partly supported in regions that are not covered by the source is considered to be particularly challenging yet more realistic (Cortes et al., 2010). We show that our method guarantees small loss independent of source-target relatedness.

The theory of active learning has also received a lot of attention in recent years (Dasgupta, 2004; Balcan et al., 2007; 2009; Hanneke, 2011). See (Dasgupta, 2011) for a survey on the main directions. However, the main goal of incorporating active queries in all these works is to learn a classifier with low error while using fewer labels. In contrast, we focus on a different benefit of active queries and formally establish that being active is also useful to adapt to

changing environments.

Nearest neighbor methods have been studied for decades (Cover & Hart, 1967; Stone, 1977; Kulkarni & Posner, 1995). While the locality of the prediction rule makes them highly flexible predictors, nearest neighbor methods suffer from a lack scalability to high dimensional data. However, there has recently been renewed interest in these methods and ways to overcome the curse of dimensionality both statistically (Kpotufe, 2011; Chaudhuri & Dasgupta, 2014) and computationally (Dasgupta & Sinha, 2013; Ram et al., 2012; Ram & Gray, 2013). Selective sampling for nearest neighbor classification has been shown to be consistent under certain conditions on the querying rule (Dasgupta, 2012); however, this work considers a data stream that comes from a fixed distribution. A 1-nearest neighbor algorithm has been analyzed under covariate shift (Ben-David & Urner, 2014); however, in contrast to our work, that study assumes a lower bound on a weight ratio between source and target. In our work, we argue that the flexibility of nearest neighbor methods can be exploited for adapting to changing environments; particularly so for choosing where to query for labels by detecting areas of the target task that are not well covered by the source.

1.3. Notation

Let $(\mathcal{X}, \rho)$ be a separable metric space. We let $B_r(x)$ denote the closed ball of radius r around x. We let $\mathrm{N}_\epsilon(\mathcal{X}, \rho)$ denote the *ϵ-covering-number* of the metric space, that is, the minimum number of subsets $C \subseteq \mathcal{X}$ of diameter at most ϵ that cover the space $\mathcal{X}$. We consider binary classification tasks, where P_S and P_T denote *source* and *target distributions* over $\mathcal{X} \times \{0,1\}$ and D_S and D_T denote their respective marginal distributions over $\mathcal{X}$. Further, $\mathcal{X}_S$ and $\mathcal{X}_T$ denote the *support* of D_S and D_T respectively. That is, for $I \in \{S,T\}$, we have $\mathcal{X}_I := \{x \in \mathcal{X} : D_I(B_r(x)) > 0 \text{ for all } r > 0\}$. We use the notation S and T for i.i.d. samples from P_S and D_T, respectively, and let $|S| = m_S$, $|T| = m_T$, and $m = m_S + m_T$. We let $\widehat{S}, \widehat{T}$ denote the empirical distributions according to S and T.

We work in the *covariate shift* setting, in which the regression function $\eta(x) = \mathbb{P}[y = 1|x]$ is the same for both source and target distributions.

For any finite $A \subseteq \mathcal{X}$ and $x \in \mathcal{X}$, the notation $x_1(x,A), \ldots, x_{|A|}(x,A)$ gives an ordering of the elements of A such that $\rho(x_1(x,A),x) \leq \rho(x_2(x,A),x) \leq \cdots \leq \rho(x_{|A|}(x,A),x)$. If A is a labeled sequence of domain points, $A = ((x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m))$, then we use the same notation for the labels (that is $y_i(x,A)$ denotes the label of the i-th nearest point to x in A). We use the notation $k(x,A) = \{x_1(x,A),\ldots,x_k(x,A)\}$ to denote the set of the k nearest neighbors of x in A.

Algorithm 1 ANDA: Active NN Domain Adaptation

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

We are interested in bounding the target loss of a k-nearest neighbor classifier. For a labeled sequence $A = ((x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m))$ we let h^k_A denote the *k-NN classifier* on A: $h^k_A(x) := \mathbf{1}\left[\frac{1}{k}\Sigma_{i=1}^k y_i(x,A) \geq \frac{1}{2}\right]$, where $\mathbf{1}[\cdot]$ is the indicator function. We denote the *Bayes classifier* by $h^*(x) = \mathbf{1}[\eta(x) \geq 1/2]$ and the *target loss* of a classifier $h : \mathcal{X} \to \{0,1\}$ by $\mathcal{L}_T(h) = \mathbb{P}_{(x,y)\sim P_T}[y \neq h(x)]$. For a subset $A \subseteq \mathcal{X}$ that is measurable both with respect to D_S and D_T and satisfies $D_T(A) > 0$, we define the *weight ratio* of A as $\beta(A) := D_S(A)/D_T(A)$. For a collection of subsets $\mathcal{B} \subseteq 2^{\mathcal{X}}$ (for example all balls in $(\mathcal{X},\rho)$), we let $d_{\mathsf{VC}}(\mathcal{B})$ denote its VC-dimension.

2. The Algorithm

In brief, our algorithm receives a labeled sample S (from the source distribution), an unlabeled sample T (from the target distribution), and two parameters k and k'. It then chooses a subset $T^l \subset T$ to be labeled, queries the labels of points in T^l, and outputs a k-NN predictor on $S \cup T^l$ (see Algorithm 1). The subset T^l is chosen so that the resulting labeled set $S \cup T^l$ is a (k,k')-NN-*cover* for the target (unlabeled) sample T.

Definition ((k,k')-NN-cover)**.** Let $T \subseteq \mathcal{X}$ be a set of elements in a metric space $(\mathcal{X},\rho)$ and let $k,k' \in \mathbb{N}$ with $k \leq k'$. A set $R \subseteq \mathcal{X}$ 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 $|k'(x, T \cup R) \cap R| \geq k$ (or both).

Our loss bound in Section 3 (Theorem 1) holds whenever $T^l \cup S$ is some (k,k')-NN-cover of T. Algorithm 2 provides a simple strategy to find such a cover: add to T^l all points whose k' nearest neighbors among $S \cup T$ include fewer than k source examples. It is easy to see that this will always result in a (k,k')-NN-cover of T. Furthermore, this approach has a query safety property: the set T^l produced by Algorithm 2 satisfies $T^l \cap Q = \emptyset$ where $Q = \{x \in T : |k'(x, S \cup T) \cap S| \geq k\}$ is the set of target examples that have k source neighbors among their k' nearest neighbors in $S \cup T$. In other words, Algorithm 2 will not query the label of any target example in regions with sufficiently many labeled source examples nearby, a property used in the query bound of Theorem 2.

2.1. Finding a Small (k,k')-NN-cover

Algorithm 2 Safe: Find a (k, k')-NN-cover

input Labeled set S, unlabeled set T, parameters k, k'
return $\{x \in T : |k'(x, S \cup T) \cap S| < k\}$

Algorithm 3 EMMA: Efficient multiset multicover approximation for finding a small (k, k')-NN-cover

input Labeled set S, unlabeled set T, parameters k, k'
$T^l \leftarrow \emptyset$
for all $x \in T$ **do**
  $r_x \leftarrow \max(0, k - k'(x, T \cup S) \cap S)$
  $n_x \leftarrow |\{x' \in T : r_{x'} > 0 \wedge x \in k'(x', S \cup T)\}|$
while $\{x \in T : r_x > 0\} \neq \emptyset$ **do**
  $T^l \leftarrow T^l \cup \{\text{argmax}_{x \in T \setminus T^l} r_x + n_x\}$
  for all $x \in T$ **do**
    $r_x \leftarrow \max(0, k - k'(x, T \cup S) \cap (S \cup T^l))$
    $n_x \leftarrow |\{x' \in T \setminus T^l : r_{x'} > 0 \wedge x \in k'(x', S \cup T)\}|$
return T^l

In order to make as few label queries as possible, we would like to find the smallest subset T^l of T to be labeled such that $T^l \cup S$ is a (k, k')-NN-cover of T. This problem is a special case of MINIMUM MULTISET MULTICOVER, a generalization of the well-known NP-hard MINIMUM SET COVER problem (see (Rajagopalan & Vazirani, 1993) and Chapter 13.2 in (Vazirani, 2001)).

Definition (MINIMUM MULTISET MULTICOVER)**.** Given a universe U of n elements, a collection of multisets $\mathcal{S}$, and a coverage requirement r_e for each element $e \in U$, we say that a multiset $S \in \mathcal{S}$ covers element e once for each copy of e appearing in S. The goal is to find the minimum cardinality set $\mathcal{C} \subseteq \mathcal{S}$ such that every element $e \in U$ is covered at least r_e times by the multisets in $\mathcal{C}$.

We can phrase the problem of finding the smallest T^l such that $T^l \cup S$ is a (k, k')-NN-cover of T as a MINIMUM MULTISET MULTICOVER problem as follows. Let $U = T$ and set the coverage requirements as $r_x = \max(0, k - |k'(x, S \cup T) \cap S|)$ for each $x \in T$. The collection $\mathcal{S}$ contains a multiset S_x for each $x \in T$, where S_x contains k copies of x and one copy of each element in $\{x' \in T : x \in k'(x', S \cup T)\}$. A minimum multiset multicover of this is also a minimum (k, k')-NN-cover and vice versa.

While MINIMUM MULTISET MULTICOVER is NP-hard to solve exactly, a greedy algorithm efficiently provides an approximate solution (see Section 2.1.1). Algorithm 3 formalizes this as an ANDA subroutine called EMMA for finding a small (k, k')-NN-cover. In the language of (k, k')-NN-covers, in each round EMMA computes the helpfulness of each $x \in T$ in two parts. The remaining coverage requirement r_x is the number of times x would cover itself if added to T^l (that is, the savings from not having to use r_x additional neighbors of x), and the total neighbor coverage n_x is the number of times x would cover its neighbors if added to T^l. EMMA then selects the point x with the largest sum $r_x + n_x$ among all points in T that have not yet been added to T^l.

In its most basic form, EMMA does not have the same query safety property enjoyed by Safe because the greedy strategy may elect to query labels of target examples that were already fully covered by source examples. We can ensure that an intelligent query strategy like EMMA still has the desired query safety property by first running Safe and then passing the resulting set T_{safe} to EMMA as its unlabeled sample. We call the resulting strategy for finding a (k, k')-NN-cover Safe-EMMA.

2.1.1. APPROXIMATION GUARANTEES

MINIMUM MULTISET MULTICOVER is known to remain NP-hard even when the multisets in $\mathcal{S}$ are small. However, a small upper bound b on the maximum size of any multiset in $\mathcal{S}$ can make the problem much easier to approximate. Specifically, the greedy algorithm has an approximation factor of H_b, the b-th harmonic number (Rajagopalan & Vazirani, 1993). This is known to be essentially optimal under standard hardness assumptions.

In our setting, the size of the largest multiset is determined by the point $x \in T$ with the largest number of points in $S \cup T$ having x as one of their k' nearest neighbors. In general metric spaces this can be up to $m = m_S + m_T$, resulting in a multiset of size $m + k$ and an approximation factor of $H_{m+k} = O(\log m)$. However, in spaces with doubling-dimension γ, it is known that $b \leq k' 4^\gamma \log_{3/2}(2L/S)$ where L and S are respectively the longest and shortest distances between any two points in T (Zhao & Teng, 2007).

3. Performance Guarantees

In this section, we analyze the expected loss of the output classifier of ANDA as well as its querying behavior. The bound in Section 3.1 on the loss holds for ANDA with any of the sub-procedures presented in Section 2. To simplify the presentation we use ANDA as a placeholder for any of ANDA-Safe, ANDA-EMMA and ANDA-Safe-EMMA. The bounds on the number of queries in Section 3.3 hold for ANDA-Safe and ANDA-Safe-EMMA, which we group under the placeholder ANDA-S.

3.1. Bounding the Loss

We start with a finite sample bound under the assumption that the regression function η satisfies a λ-Lipschitz condition. That is, we have $|\eta(x) - \eta(x')| \leq \lambda\rho(x, x')$ for all $x, x' \in \mathcal{X}_S \cup \mathcal{X}_T$.

Our bound on the expected loss in Theorem 1 is proven

using standard techniques for nearest neighbor analysis. However, since our algorithm does not predict with a fully labeled sample from the target distribution (possibly very few or even none of the target generated examples get actually labeled and the prediction is mainly based on source generated examples), we need to ensure that the set of labeled examples still sufficiently covers the target task. The following lemma serves this purpose. It bounds the distance of an arbitrary domain point x to its k-th nearest *labeled point* in terms of its distance to its k'-th nearest *target sample point*. Note that the bound in the lemma is easy to see for points in T. However, we need it for arbitrary (test-) points in the domain.

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

Proof. Let $x \in \mathcal{X}$. If the set $k'(x, T)$ of the k' nearest neighbors of x in T contains k points from R, we are done (in this case we actually have $\rho(x, x_k(x, R)) \leq \rho(x, x_{k'}(x, T))$). Otherwise, let $x' \in k'(x, T) \setminus R$ be one of these points that is not in R. Since R is a (k, k')-NN-cover for T, and $x' \in T$, the set of the k' nearest neighbors of x' in $R \cup T$ contains k elements from R.

Let x'' be any of these k elements, that is $x'' \in R \cap k'(x', R \cup T)$. Note that $\rho(x', x'') \leq 2\rho(x, x_{k'}(x, T))$ since x' is among the k' nearest neighbors of x and x'' is among the k' nearest neighbors of x' in $R \cup T$. Thus, we have

$$\begin{aligned}\rho(x, x'') &\leq \rho(x, x') + \rho(x', x'') \\ &\leq \rho(x, x_{k'}(x, T)) + 2\rho(x, x_{k'}(x, T)) \\ &= 3\rho(x, x_{k'}(x, T)).\end{aligned}$$

□

This lemma allows us to establish the finite sample guarantee on the expected loss of the classifier output by ANDA. Note that the guarantee in the theorem below is independent of the size and the generating process of S (except for the labels being generated according to η), while possibly (if S covers the target sufficiently) only few target points are queried for labels. Recall that $\mathrm{N}_\epsilon(\mathcal{X}_T, \rho)$ denotes the ϵ-covering number of the target support.

Theorem 1. *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,*

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

The proof (see supplementary material, Section 1) incorporates our bound on the distance to the k nearest labeled points of Lemma 1 into a standard technique for nearest neighbor analysis (as in (Shalev-Shwartz & Ben-David, 2014)). The key to the guarantee being the bound in Lemma 1, one could obtain analogous generalization bounds under relaxed assumptions for which nearest neighbor classification can be shown to succeed (see, e.g. (Chaudhuri & Dasgupta, 2014) for a discussion on such). Similarly, one could obtain bounds for other settings, such as multi-class classification and regression.

3.2. Consistency

We show that ANDA is consistent in a slightly more general setting, namely if the regression function is *uniformly continuous* and the $\mathrm{N}_\epsilon(\mathcal{X}_T, \rho)$ are finite. Note that this is the case, for example, if $(\mathcal{X}, \rho)$ is compact and η is continuous. Recall that a function $\eta : \mathcal{X} \to \mathbb{R}$ is *uniformly continuous* if for every $\gamma > 0$ there exists a δ such that for all $x, x' \in \mathcal{X}$, $\rho(x, x') \leq \delta \Rightarrow |\eta(x) - \eta(x')| \leq \gamma$. The proof is located in the supplementary material, Section 2.

Corollary 1. *Let $(\mathcal{X}, \rho)$ be a metric space, and let $\mathcal{P}(\mathcal{X}, \rho)$ denote the class of distributions over $\mathcal{X} \times \{0, 1\}$ with uniformly continuous regression functions. Let $(k_i)_{i \in \mathbb{N}}$, $(k'_i)_{i \in \mathbb{N}}$ and $(m_i)_{i \in \mathbb{N}}$ be non-decreasing sequences of natural numbers with $k'_i \geq k_i$ for all i, and $k_i \to \infty, k'_i \to \infty, m_i \to \infty$ and $(k'_i / m_i) \to 0$ as $i \to \infty$. For each $i \in \mathbb{N}$, let $S_i \in (\mathcal{X} \times \{0, 1\})^{n_i}$ be a sequence of labeled domain points. Then for any distribution $P_T \in \mathcal{P}(\mathcal{X}, \rho)$ with finite covering numbers $\mathrm{N}_\epsilon(\mathcal{X}_T, \rho)$, we have*

$$\lim_{i \to \infty} \mathop{\mathbb{E}}_{T \sim P_T^{m_i}} [\mathcal{L}_T(\mathrm{ANDA}(S_i, T, k_i, k'_i))] = \mathcal{L}_T(h^*).$$

3.3. Bounding the Number of Queries

In this section, we show that our algorithm automatically adapts the number of label queries to the similarity of source and target task. First, we now provide a finite sample bound that implies that with a sufficiently large source sample, with high probability, ANDA-S does not query at all in areas where the weight ratio of balls is bounded from below; i.e. it only queries where it is "needed." In our analysis, we employ a lemma by (Kpotufe, 2011), which follows from VC-theory (Vapnik & Chervonenkis, 1971).

Lemma 2 (Lemma 1 in (Kpotufe, 2011))**.** *Let $\mathcal{B}$ denote the class of balls in $(\mathcal{X}, \rho)$, and let D be a distribution over $\mathcal{X}$. Let $0 < \delta < 1$, and define $\alpha_n = (d_{\mathsf{VC}}(\mathcal{B}) \ln(2n) + \ln(6/\delta))/n$. The following holds with probability at least $1 - \delta$ (over a sample T of size n drawn i.i.d. from D) for all balls $B \in \mathcal{B}$: if $a \geq \alpha_n$, then $\widehat{T}(B) \geq 3a$ implies $D(B) \geq a$ and $D(B) \geq 3a$ implies $\widehat{T}(B) \geq a$.*

With this, we now prove our query bound. We let $B_{k,T}(x)$ denote the smallest ball around x that contains the k nearest

neighbors of x in T, and $\mathcal{B}$ the class of all balls in $(\mathcal{X}, \rho)$. Recall that $\beta(B) = D_S(B)/D_T(B)$ is the weight ratio.

Theorem 2. *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 $k \geq 9\,(d_{\mathsf{VC}}(\mathcal{B})\ln(2m_T) + \ln(6/\delta))$. Let the source sample size satisfy*

$$m_S \geq \frac{72\,\ln(6/\delta) m_T}{C\,w} \ln\left(\frac{9\,m_T}{C\,w}\right)$$

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 $\beta(B_{Ck,T}(x)) > w$.*

Proof. Since $k \geq 9\,(d_{\mathsf{VC}}(\mathcal{B})\ln(2m_T) + \ln(6/\delta))$, we have $d_{\mathsf{VC}}(\mathcal{B})/k < 1$. Thus, we get

$$m_S \geq \max\left\{8\left(\frac{9\,d_{\mathsf{VC}}(\mathcal{B})\,m_T}{C\,k\,w}\right)\ln\left(\frac{9\,d_{\mathsf{VC}}(\mathcal{B})\,m_T}{C\,k\,w}\right), \frac{18\,\ln(6/\delta)\,m_T}{C\,k\,w}, \frac{9\,m_T}{C\,w}\right\},$$

Note that $m_S \geq 8\left(\frac{9\,d_{\mathsf{VC}}(\mathcal{B})\,m_T}{C\,k\,w}\right)\ln\left(\frac{9\,d_{\mathsf{VC}}(\mathcal{B})\,m_T}{C\,k\,w}\right)$ implies that $m_S \geq 2\left(\frac{9\,d_{\mathsf{VC}}(\mathcal{B})\,m_T}{C\,k\,w}\right)\ln(2m_S)$, and together with the second lower bound (in the max) on m_S, this yields

$$m_S \frac{C\,k\,w}{3\,m_T} \;\geq\; 3(d_{\mathsf{VC}}(\mathcal{B})\ln(2m_S) + \ln(6/\delta)). \qquad (1)$$

We now assume that S and T are so that the implications in Lemma 2 are valid (this holds with probability at least $1 - 2\delta$ over the samples S and T). Let $x \in T$ be such that $\beta(B_{Ck,T}(x)) > w$. By definition of the ball $B_{Ck,T}(x)$, we have $\widehat{T}(B_{Ck,T}(x)) = \frac{Ck}{m_T}$, and by our choice of k, therefore

$$\widehat{T}(B_{Ck,T}(x)) \;=\; \frac{C\,k}{m_T} \;\geq\; \frac{C\,9\,(d_{\mathsf{VC}}(\mathcal{B})\ln 2m_T + \ln 6/\delta)}{m_T}.$$

Now Lemma 2 implies that $D_T(B_{Ck,T}(x)) \;\geq\; \frac{C\,k}{3\,m_T}$, so the condition on the weight ratio of this ball now yields

$$\begin{aligned} D_S(B_{Ck,T}(x)) \;&\geq\; \frac{C\,k\,w}{3\,m_T} \;=\; m_S \frac{C\,k\,w}{3\,m_T\,m_S} \\ &\geq\; 3\left(\frac{d_{\mathsf{VC}}(\mathcal{B})\ln(2m_S) + \ln(6/\delta)}{m_S}\right), \end{aligned}$$

where the last inequality follows from Equation (1). Now, Lemma 2, together with $m_S \geq \frac{9\,m_T}{C\,w}$ (the third term in the max), implies $\widehat{S}(B_{Ck,T}(x)) \;\geq\; \frac{C\,k\,w}{9\,m_T} \;\geq\; \frac{k}{m_S}$. This means that $B_{Ck,T}(x)$ contains k examples from the source, which implies that among the $k' = Ck + k$ nearest sample points (in $S \cup T$) there are k source examples, and therefore x will not be queried by ANDA-S. □

Theorem 2 provides a desirable guarantee for the "lucky" case: It implies that if the source and target distributions happen to be identical or very similar, then, given that ANDA-S is provided with a sufficiently large source sample, it will not make any label queries at all. More importantly, the theorem shows that, independent of an *overall* source/target relatedness measure, the querying of ANDA-S adapts automatically to a *local* relatedness measure in the form of weight ratios of balls around target sample points. ANDA-S queries only where it is necessary to compensate for insufficient source coverage.

3.4. Query Consistency

Extending the proof technique of Theorem 2, we get a "query-consistency" result under the assumption that D_S and D_T have continuous density functions. In the limit of large source samples, ANDA-S will, with high probability, not make any queries in the source support. The proof is in the supplementary material, Section 3.

Theorem 3. *Let D_S and D_T have continuous density functions. Let $\delta > 0$, $C > 1$, and let m_T, k and k' satisfy the conditions of Theorem 2. Then, there exists a (sufficiently large) source sample size M_S such that with probability at least $(1 - 3\delta)$ over source samples of size $m_S \geq M_S$ and target samples of size m_T,* ANDA-S *will not make any label queries in the source support.*

Together with Corollary 1 this shows that, for increasing target sample sizes, the expected loss of the output of ANDA-S converges to the Bayes optimal and, with high probability over increasing source samples, ANDA-S will not query target sample points in the source support.

4. Experiments

Our experiments on synthetic data illustrate ANDA's adaptation ability and show that its classification performance compares favorably with baseline passive nearest neighbors. Experiments on challenging image classification tasks show that ANDA is a good candidate for correcting dataset bias. We discuss the results in relation to our theory.

4.1. Synthetic Data

The source marginal D_S was taken to be the uniform distribution over $[-1, 0.5]^2$ and the target marginal D_T was set to uniform over $[-0.75, 1]^2$. This ensures enough source/target overlap so the source data is helpful in learning the target task but not sufficient to learn well. The regression function chosen for both tasks was $\eta(x_1, x_2) = (1/2)(1 - (\sin(2\pi x_1)\sin(2\pi x_2))^{1/6})$ for $(x_1, x_2) \in \mathbb{R}^2$. This creates a 4×4 checkerboard of mostly-positively and mostly-negatively labeled regions with noise on the boundaries where η crosses $1/2$. Training samples from this set-

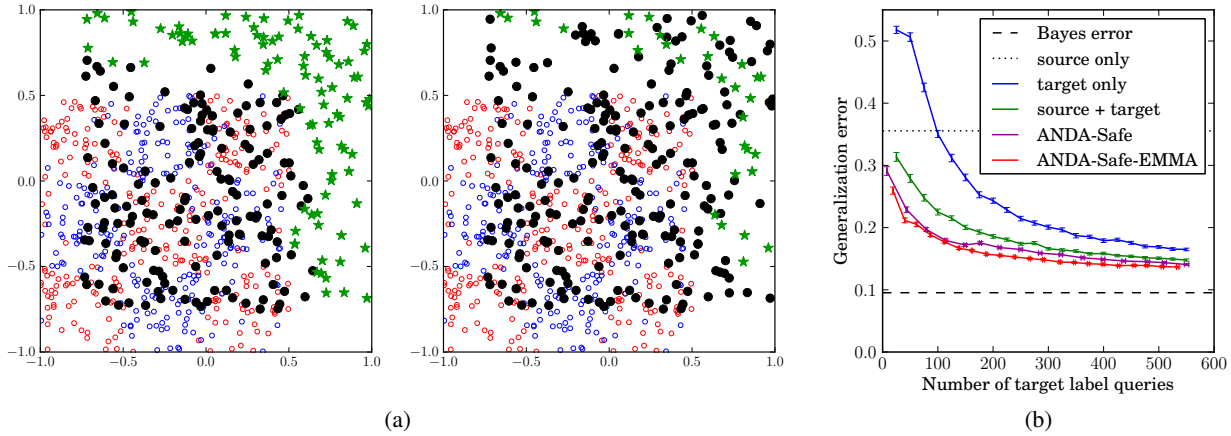


Figure 1. (a) Visualization of synthetic data and query strategies for ANDA-Safe (left) and ANDA-Safe-EMMA (right). Red and blue circles represent labeled source examples, black circles represent unqueried target examples, and green stars represent queried target examples. (b) Experimental results on synthetic data. Error bars represent two standard errors, or roughly a 95% confidence interval.

ting are pictured in Figure 1(a) along with query locations. Notice that queries are almost never made inside the source support, as our theory would suggest.

The baseline algorithms we compare against are the following. The "source only" algorithm predicts according to a k-NN classifier built on a source sample alone. The "target only" algorithm creates a k-NN classifier on a random sample from the target, and "source + target" does the same but includes labeled data from a source sample as well.

We compare the generalization error of ANDA-Safe-EMMA and ANDA-Safe against these baselines across a range of unlabeled target sample sizes. Since the number of queries made by both ANDA-Safe-EMMA and ANDA-Safe increases with target sample size, this generates a range of query counts for the active algorithms. The baseline algorithms were given labeled target samples of sizes in the same range as these query counts. For all algorithms and target sample sizes we fixed $m_S = 3200$, $k = 7$, and $k' = 21$. Figure 1(b) shows the resulting generalization error (averaged over 100 independent trials) for each algorithm as a function of the number of target labels used.

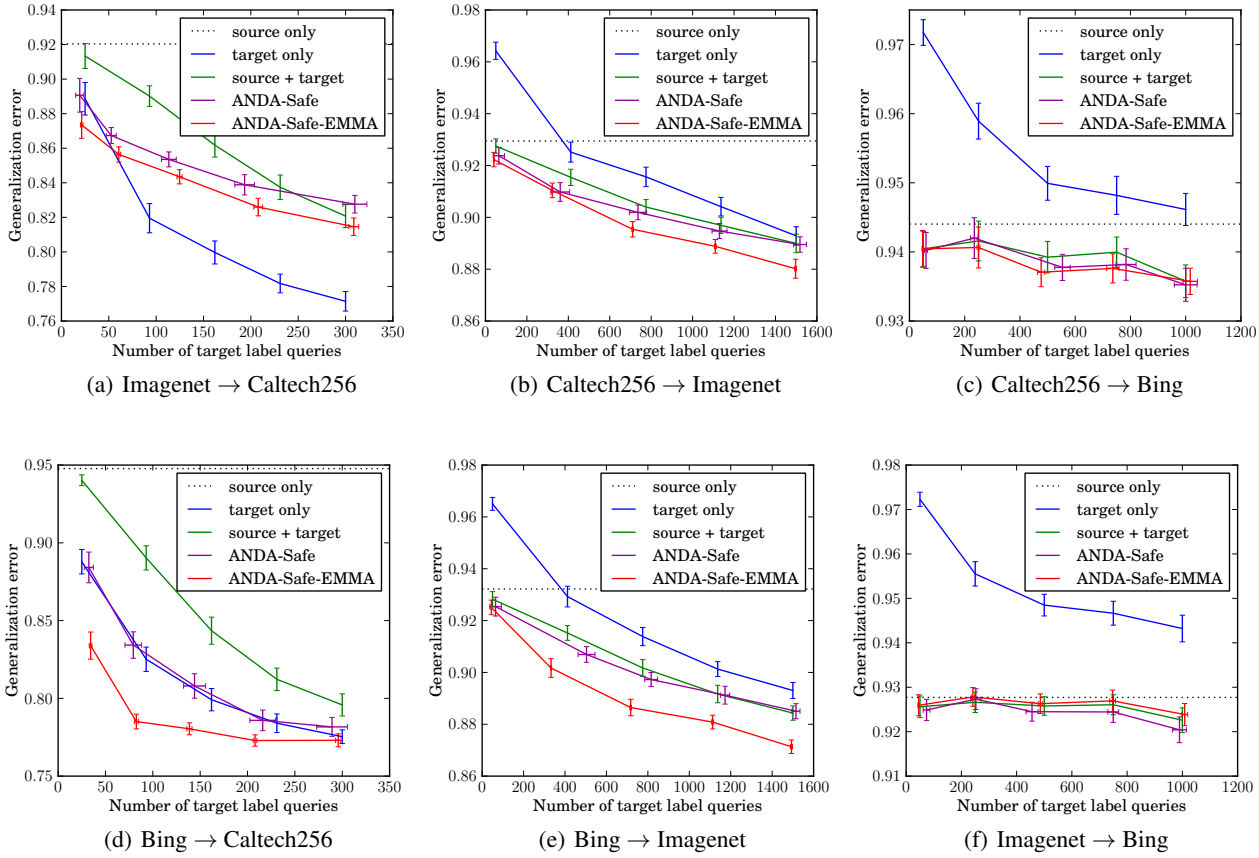
Both active algorithms perform significantly better than the passive baselines in terms of the error they achieve per target label query. ANDA-Safe-EMMA also outperforms ANDA-Safe, since (as shown in Figure 1(a)) achieves full coverage of the target region with many fewer queries.

4.2. Image Classification

A major problem in building robust image classifiers is that the source of training images is often not the same as the source of images on which the classifier is expected to perform. This leads to *dataset bias*, which requires domain adaptation to correct. Tommasi & Tuytelaars (2014) aligned and preprocessed several image datasets that provide a way of comparing domain adaptation methods on this problem. Even though these datasets are unlikely to satisfy covariate shift exactly, we compare ANDA with baseline nearest neighbor classifiers to show that ANDA provides a partial solution to the dataset bias problem.

The task is to classify images according to the object in the image. We use the dense setup which contains four datasets (representing different domains) and 40 object classes. SIFT features for each image were precomputed and grouped into a bag-of-words representation with a 1000-word vocabulary. Despite the high dimensionality, we find that nearest neighbor methods work well on these datasets without further dimensionality reduction. Of the four datasets (Caltech256, Imagenet, Bing, and SUN) we chose not to use SUN because the differences in how data was labeled result in a clear violation of covariate shift.

Each of the other three datasets was used twice as source data and twice as the target. For each of the six source/target combinations, we compared the same algorithms described in Section 4.1 and used the same method for generating a range of query counts. For all algorithms and target sample sizes we fixed $m_S = 2000$, $k = 25$, and $k' = 75$. Figure 2 shows the resulting generalization error (estimated from test sets of 1000 examples and averaged over 50 independent trials) for each algorithm as a function of the number of target labels used. The error values reported here are on the same order as those in the results of Tommasi & Tuytelaars (2014), but since different sample



(a) Imagenet → Caltech256 (b) Caltech256 → Imagenet (c) Caltech256 → Bing

(d) Bing → Caltech256 (e) Bing → Imagenet (f) Imagenet → Bing

Figure 2. Results on image classification task. Each plot caption is of the form *source* → *target*. Error bars represent two standard errors.

sizes were used, they cannot be directly compared[1].

Overall we find that our methods (especially ANDA-Safe-EMMA) successfully correct for dataset bias in image classification, also showing that ANDA is robust to small violations of our theory's assumptions. For all 6 pairs of datasets, ANDA-Safe-EMMA performs better than using source data alone (adding in target examples always helps). Even more encouraging, on 5 of the 6 pairs, it performs better than the target-only baseline (indicating that having source examples allows us to make more efficient use of target labels) and on 4 of the 6 it outperforms the passive source + target baseline (and never performs worse).

When Bing is the target (Figures 2(c) and 2(f)), neither active algorithm performs better (or worse) than the passive baseline. Bing was previously known to be noisier than the other two datasets (Tommasi & Tuytelaars, 2014), and further evidence of this can be found in the observation that the source-only baselines (for both Caltech256 and Imagenet) perform better than Bing's target-only baseline. This means the target queries from Bing are generally less informative than source examples, regardless of where the queries are made, resulting in all the source/target combination methods performing equally well.

When Caltech256 is the target (Figures 2(a) and 2(d)), the target-only baseline outperforms the other methods for high enough query counts. This is likely because Caltech256 is less noisy than the other datasets, so the noisy source data is helpful in the absence of target data but harmful when enough target data is available. Notice that for both of these cases, ANDA-Safe-EMMA has the best accuracy at small query counts, exemplifying its efficiency at making use of labels when it only makes a few queries.

Finally, we ran experiments with source and target sets sampled from the same dataset (using the same parameters as above). ANDA does not make any label queries once $m_S \geq m_T$. This confirms a desirable property predicted by our theory: ANDA will automatically detect when to rely on source data alone and not waste label queries.

[1]Note that since there are 40 classes, guessing labels uniformly at random results in a generalization error of 97.5%.

Acknowledgments

We would like to thank Nina Balcan for her vision and inspiration and for her support throughout the course of this project. Parts of this work were done while the second author was a postdoctoral fellow at Georgia Institute of Technology and at Carnegie Mellon University. This work was supported in part by NSF grants CCF-1451177, CCF-1101283, CCF-1422910, ONR grant N00014-09-1-0751, AFOSR grant FA9550-09-1-0538, and a Microsoft Faculty Fellowship.

References

Balcan, Maria-Florina, Broder, Andrei, and Zhang, Tong. Margin-based active learning. In *COLT*, 2007.

Balcan, Maria-Florina, Beygelzimer, Alina, and Langford, John. Agnostic active learning. *J. Comput. Syst. Sci.*, 75 (1), 2009.

Ben-David, Shai and Urner, Ruth. Domain adaptation-can quantity compensate for quality? *Ann. Math. Artif. Intell.*, 70(3):185–202, 2014.

Ben-David, Shai, Blitzer, John, Crammer, Koby, and Pereira, Fernando. Analysis of representations for domain adaptation. In *NIPS*, 2006.

Chattopadhyay, Rita, Fan, Wei, Davidson, Ian, Panchanathan, Sethuraman, and Ye, Jieping. Joint transfer and batch-mode active learning. In *ICML*, 2013a.

Chattopadhyay, Rita, Wang, Zheng, Fan, Wei, Davidson, Ian, Panchanathan, Sethuraman, and Ye, Jieping. Batch mode active sampling based on marginal probability distribution matching. *TKDD*, 7(3):13, 2013b.

Chaudhuri, Kamalika and Dasgupta, Sanjoy. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pp. 3437–3445, 2014.

Cortes, Corinna, Mansour, Yishay, and Mohri, Mehryar. Learning bounds for importance weighting. In *NIPS*, 2010.

Cover, Thomas M. and Hart, Peter E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Dasgupta, Sanjoy. Analysis of a greedy active learning strategy. In *NIPS*, 2004.

Dasgupta, Sanjoy. Two faces of active learning. *Theor. Comput. Sci.*, 412(19):1767–1781, 2011.

Dasgupta, Sanjoy. Consistency of nearest neighbor classification under selective sampling. In *COLT*, 2012.

Dasgupta, Sanjoy and Sinha, Kaushik. Randomized partition trees for exact nearest neighbor search. In *COLT*, 2013.

Hanneke, Steve. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

Kpotufe, Samory. k-NN regression adapts to local intrinsic dimension. In *NIPS*, 2011.

Kulkarni, Sanjeev R. and Posner, S. E. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41 (4):1028–1039, 1995.

Mansour, Yishay, Mohri, Mehryar, and Rostamizadeh, Afshin. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

Pan, Sinno Jialin and Yang, Qiang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. ISSN 1041-4347.

Rajagopalan, Sridhar and Vazirani, Vijay V. Primal-dual RNC approximation algorithms for (multi)-set (multi)-cover and covering integer programs. In *FOCS*, 1993.

Ram, Parikshit and Gray, Alexander G. Which space partitioning tree to use for search? In *NIPS*, 2013.

Ram, Parikshit, Lee, Dongryeol, and Gray, Alexander G. Nearest-neighbor search on a time budget via max-margin trees. In *SDM*, 2012.

Saha, Avishek, Rai, Piyush, III, Hal Daumé, Venkatasubramanian, Suresh, and DuVall, Scott L. Active supervised domain adaptation. In *ECML/PKDD*, 2011.

Settles, Burr. Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66, 2010.

Shalev-Shwartz, Shai and Ben-David, Shai. *Understanding Machine Learning*. Cambridge University Press, 2014.

Shi, Yuan and Sha, Fei. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *ICML*, 2012.

Stone, Charles J. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 07 1977.

Sugiyama, Masashi, Suzuki, Taiji, Nakajima, Shinichi, Kashima, Hisashi, von Bünau, Paul, and Kawanabe, Motoaki. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Tommasi, Tatiana and Tuytelaars, Tinne. A testbed for cross-dataset analysis. In *TASK-CV Workshop at ECCV*, 2014.

Vapnik, Vladimir N. and Chervonenkis, Alexey J. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Vazirani, Vijay. *Approximation Algorithms*. Springer, 2001.

Zhao, Yingchao and Teng, Shang-Hua. Combinatorial and spectral aspects of nearest neighbor graphs in doubling dimensional and nearly-euclidean spaces. In *TAMC*, 2007.