Domain Adaptation using Nearest Neighbors

Ruth Urner, Shai Ben-David
David R. Cheriton
School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1
CANADA
{rurner,shai}@cs.uwaterloo.ca

Shai Shalev-Shwartz
School of Computer Science & Engineering
The Hebrew University of Jerusalem
Givat Ram, Jerusalem 91904
ISRAEL
shais@cs.huji.ac.il

Abstract

The domain adaptation problem in machine learning occurs when the test data generating distribution differs from the one that generates the training data. We study simple assumptions about the relationship between the two distributions that suffice for domain adaptation learning to succeed. We propose two types of assumptions. (1) The weight ratio assumption in which we bound the ratio of the probability weights of a restricted class of subsets of the domain between the train/test marginal (unlabeled) distributions. (2) A novel probabilistic relaxation of the classic Lipschitzness condition applied on the underlying labeling function. We analyze domain adaptation with the nearest neighbor algorithm using these two assumptions and provide finite sample guarantees. We augment our positive learnability results with lower bounds, showing that none of these assumptions suffices on its own. We also discuss the implications of our results for *proper* domain adaptation learning, where the learner is required to output a predictor from some pre-determined class. We propose a learning algorithm that utilizes the unlabeled target training sample in an essential way—we prove that for proper learning, any algorithm that has no access to such samples, will produce a significantly larger error than ours. To the best of our knowledge, this is the first result proving that the use of target-generated unlabeled samples can be utilized to achieve a clear performance advantage.

1 Introduction

Much of the theoretical analysis of machine learning is focused on the case when the training and test data are generated by the *same* underlying distribution. While this may sometimes be a good approximation of reality, in many practical tasks this assumption cannot be justified. For example, when building email spam detectors, the training data are emails received at some address but the filter is to be applied to emails sent to a different user. Even when the training emails are collected at the same account to which learned filter will be applied, it may be that the email-generating distribution changes over time, thus creating some discrepancy between the training and test data distributions.

Nevertheless, this is not an "all-or-nothing" situation. While the training and test distributions may not be completely identical, they are often quite similar. Furthermore, in some such tasks, unlabeled examples, generated by the distribution governing the target domain, may be also available to the learner. We address the following high-level questions:

- What conditions, mainly on the relationship between the source training and target test distributions, allow DA to succeed? Which assumptions suffice to provide performance guarantees on the success of DA algorithms?
- Which algorithmic paradigms are likely to perform well under such given relatedness assumptions?

The learning model that we analyze here has two data-generating distributions: a source distribution and a target distribution. Both generate labeled examples. The DA learner has access to an i.i.d. *labeled* sample from the source distribution, and to an i.i.d. *unlabeled* sample from the target distribution. The learner is expected to output a predictor, whose success is evaluated with respect to the target distribution.

The first assumption we impose throughout the paper is that the conditional label distributions are the same for the source and target distribution. This restriction is the commonly referred to as the *covariate-shift assumption*, and is often assumed in domain adaptation analysis (Sugiyama and Mueller (2005)). Besides the covariate shift assumption we propose and analyze two types of additional assumptions:

1. Bounding the ratio of the probability weights between the two marginal (unlabeled) distributions. Previous work (e.g. Cortes et al. (2010)) assumes an upper bound on the point-wise density ratio. We propose a family of relaxations of this assumption in which we only require a bound on the ratio of source and target weights of subsets of the domain that comes from some predefined collection of subsets.

2. Taming the behavior of the conditional labeling function. We introduce a novel probabilistic relaxation of the classic Lipschitzness condition. In contrast with the common notion, the probabilistic Lipschitzness is also meaningful for binary-valued functions over connected domains.

In the main part of the paper we prove that the combination of such assumptions is sufficient to allow non-parametric domain adaptation learning and provide sample size bounds in terms of the value of the corresponding parameters. The domain adaptation algorithm that we apply in this context is based on the nearest neighbor paradigm. As part of our analysis, we derive finite sample size error bounds for the nearest neighbor algorithm, bounds that can be also applied to the usual, single data distribution, case. We augment our positive learnability results by deriving corresponding lower bounds. We prove that none of the two assumptions mentioned above suffices on its own, and that the dependence of the sample complexity on the Lipschitzness condition of the labeling function is inevitable, even when the ratio of the marginal probabilities is favorable. Our lower bounds involve a novel reduction from some basic statistical tasks.

We also discuss the setup of *proper* DA learning, where the learner is required to output a predictor from some pre-determined class. Such learners are relevant either when we assume, as a prior knowledge, that the labeling function comes from a predefined hypotheses class, or when additional requirements are imposed on the learned predictor. For example, if the output predictor must be very fast, we can restrict the learner to output a predictor from a hypotheses class that only contains fast-computable functions. For proper DA learning, we derive two positive results. First, we show that in the realizable case, when the labeling function indeed belongs to our hypotheses class, then the weight ration assumption suffices for DA learnability. Second, in the unrealizable case, we propose a DA learner that utilizes the unlabeled target training sample in an essential way—we prove that no proper learning algorithm can achieve learning rates similar to ours, without access to target-generated samples. As far as we are aware, there are no previous results showing a clear advantage obtained by the usage of target-generated unlabeled finite samples of bounded size.

References

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 442–450. 2010.

M. Sugiyama and K. Mueller. Generalization error estimation under covariate shift. In *Workshop on Information-Based Induction Sciences*, 2005.