Theoretical Computer Science 519 (2014) 103–126

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Domain adaptation and sample bias correction theory and algorithm for regression

Corinna Cortes [a,*], Mehryar Mohri [b,a]

[a] Google Research, 76 Ninth Avenue, New York, NY 10011, United States
[b] Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, United States

ARTICLE INFO

Keywords:
Machine learning
Learning theory
Domain adaptation
Optimization

ABSTRACT

We present a series of new theoretical, algorithmic, and empirical results for domain adaptation and sample bias correction in regression. We prove that the discrepancy is a distance for the squared loss when the hypothesis set is the reproducing kernel Hilbert space induced by a universal kernel such as the Gaussian kernel. We give new pointwise loss guarantees based on the discrepancy of the empirical source and target distributions for the general class of kernel-based regularization algorithms. These bounds have a simpler form than previous results and hold for a broader class of convex loss functions not necessarily differentiable, including L_q losses and the hinge loss. We also give finer bounds based on the discrepancy and a weighted feature discrepancy parameter. We extend the discrepancy minimization adaptation algorithm to the more significant case where kernels are used and show that the problem can be cast as an SDP similar to the one in the feature space. We also show that techniques from smooth optimization can be used to derive an efficient algorithm for solving such SDPs even for very high-dimensional feature spaces and large samples. We have implemented this algorithm and report the results of experiments both with artificial and real-world data sets demonstrating its benefits both for general scenario of adaptation and the more specific scenario of sample bias correction. Our results show that it can scale to large data sets of tens of thousands or more points and demonstrate its performance improvement benefits.

1. Introduction

A standard assumption in learning theory and the design of learning algorithms is that training and test points are drawn according to the same distribution. In practice, however, this assumption often does not hold. A more challenging problem of *domain adaptation* arises in a variety of applications, including natural language processing, speech processing, or computer vision [12,4,17,18,29,30,21]. This problem occurs when little or no labeled data is available from the *target domain*, but labeled data from a *source domain* somewhat similar to the target, as well as large amounts of unlabeled data from the target domain, are accessible. The domain adaptation problem then consists of using the source labeled and target unlabeled data to learn a hypothesis performing well on the target domain.

The theoretical analysis of this problem has been the topic of some recent publications. The first theoretical analysis of adaptation is due to Ben-David et al. [1] (some technical issues of that paper were later corrected by Blitzer et al. [5]). These authors gave VC-dimension bounds for binary classification based on a d_A distance between distributions that can be estimated from finite samples, and a term λ_H depending on the distributions and the hypothesis set H, which cannot be

* Corresponding author.
E-mail addresses: corinna@google.com (C. Cortes), mohri@cims.nyu.edu (M. Mohri).

0304-3975/$ – see front matter
http://dx.doi.org/10.1016/j.tcs.2013.09.027

estimated from data. The assumptions made in the analysis of adaptation were more recently discussed by Ben-David et al. [2] who presented some negative results for adaptation in the case of the zero-one loss based on a handful of examples.

In previous work [20], we introduced the notion of *discrepancy* which generalizes the d_A distance to arbitrary loss functions. We gave data-dependent Rademacher complexity bounds showing how the discrepancy can be estimated from finite samples. We then presented alternative learning bounds for adaptation based on the discrepancy. These bounds hold for a general class of loss functions, including the zero-one loss function used in classification, and depend on the optimal classifiers in the hypothesis set for the source and target distributions. They are in general not comparable to those of Ben-David et al. [1] or Blitzer et al. [5], but we showed that, under some plausible assumptions, they are superior to those of [1,5] and that in many cases the bounds of Ben-David et al. [1] or Blitzer et al. [5] have a factor of 3 of the error that can make them vacuous. Perhaps more importantly, we also gave a series of pointwise loss guarantees for the broad class of kernel-based regularization algorithms in terms of the empirical discrepancy. These bounds motivated a discrepancy minimization algorithm and we initiated the study of its properties.

Many of the previous techniques or paradigms used for adaptation and similar problems consist of reweighting the training point losses to more closely reflect those in the test distribution. The definition of the reweighting is of course crucial and varies for different techniques. A common choice consists of selecting the weight of a point x of the training sample as an estimate of the ratio $\omega(x) = P(x)/Q(x)$ where P is the target (unbiased) distribution and Q the observed source distribution, since this choice preserves the expected loss [32,10]. However, we gave an empirical and theoretical analysis of importance weighting [11] which shows that, even when using the exact ratio P/Q, such importance weighting techniques do not succeed in general, even in the simple case of two Gaussians. A critical issue we pointed out is that the weight $\omega(x)$ is unbounded in many practical cases and can become very large for a few points x of the sample that end up fully dominating the learning process, thereby resulting in a very poor performance. We also presented an analysis of the effect of an error in the estimation of the reweighting parameters on the accuracy of the hypothesis returned by the learning algorithm in [10].

Bickel et al. [3] developed a discriminative model that instead characterizes how much more likely an instance is to occur in the test sample than in the training sample. The optimization solution they describe is in general not convex but they prove it to be in the case of the exponential loss. Their method is proposed for a classification setting but can also be applied to the regression setting in a two-stage approximation. Weights are first learned by maximizing the posterior probability given all the available data. These weights can then be used in combination with any algorithm that allows for weighted examples. A somewhat different *kernel mean matching* (KMM) method was described by Huang et al. [16] in the context of kernel methods. This consists of defining the weights assigned to the training sample in a way such that the mean feature vector on the training points be as close as possible to the mean feature vector over the unlabeled points. Yu and Szepesvári recently presented an analysis of the KMM estimator [37]. This should be distinguished from an analysis of the generalization properties of KMM as an algorithm for adaptation or sample bias correction. Sugiyama et al. [34] argued that KMM does not admit a principled cross-validation method helping to select the kernel parameters and proposed instead an algorithm (KLIEP) addressing that issue. Their algorithm determines the weights based on the minimization of the relative entropy of $\omega(x)Q(x)$ and the distribution of unlabeled data over the input domain, or equivalently, the maximization of the log-likelihood of $\omega(x)Q(x)$ for the observed unlabeled data. The weights $\omega(x)$ are modeled, more specifically, as a linear combination of kernel basis functions such as Gaussians. The algorithms just mentioned do not take into account the hypothesis set used by the learning algorithm, or the loss function relevant to the problem. In contrast, we will introduce and analyze an algorithm that precisely takes both of these into consideration.

In this paper, we present a series of novel results for domain adaptation extending those of [20] and making them more significant and practically applicable.[1] Our analysis concentrates on the problem of adaptation in regression. We also consider the problem of sample bias correction, which can be viewed as special instance of adaptation.

In Section 2, we describe more formally the learning scenario of domain adaptation in regression and briefly review the definition and key properties of the discrepancy. We then present several new theoretical results in Section 3. For the squared loss, we prove that the discrepancy is a distance when the hypothesis set is the reproducing kernel Hilbert space of a universal kernel, such as a Gaussian kernel. This implies that minimizing the discrepancy to zero guarantees matching the target distribution, a result that does not hold in the case of the zero-one loss. We further give pointwise loss guarantees depending on the discrepancy of the empirical source and target distributions for the class of kernel-based regularization algorithms, including kernel ridge regression, support vector machines (SVMs), or support vector regression (SVR). These bounds have a simpler form than a previous result we presented in the specific case of the squared loss in [20] and hold for a broader class of convex loss functions not necessarily differentiable, which includes all L_q losses ($q \geqslant 1$), but also the hinge loss used in classification. We also present finer bounds in the specific case of the squared loss L_2 in terms of the discrepancy and a *weighted feature discrepancy* parameter that we define and analyze in detail.

When the magnitude of the difference between the source and target labeling functions is small on the training set, these bounds provide a strong guarantee based on the empirical discrepancy and suggest an adaptation algorithm based on empirical discrepancy minimization [20] detailed in Section 4. In Section 5, we extend the discrepancy minimization algorithm with the squared loss to the more significant case where kernels are used. We show that the problem can be cast

[1] This is an extended version of the conference paper [8] including more details and additional theoretical and empirical results.