Signal Processing 108 (2015) 272–277

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Fast communication

Active data labeling for improved classifier generalizability

Visar Berisha*, Douglas Cochran

Arizona State University, Tempe, AZ 85287, USA

ARTICLE INFO

Article history:
Received 12 June 2014
Accepted 12 September 2014
Available online 22 September 2014

Keywords:
Classification
Divergence measures
Domain adaptation
Active learning

ABSTRACT

Existing statistical learning methods perform well when evaluated on training and test data drawn from the same distribution. In practice, however, these distributions are not always the same. In this paper we derive an estimable upper bound on the test error rate that depends on a new probability distance measure between training and test distributions. Furthermore, we identify a non-parametric estimator for this distance measure that can be estimated directly from data. We show how this new probability distance measure can be used to construct algorithmic tools that improve performance. In particular, motivated by our upper bound, we propose a new active learning algorithm for domain adaptation. Comparative results confirm the efficacy of the active learning algorithm on a set of 12 speech classification tasks.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Many signal processing and machine learning tasks require training data to learn the parameters of some underlying model. Most of the theoretical analyses and practical considerations in these applications are focused on scenarios where training and test data come from the same distribution. This may be justified in a number of small, controlled circumstances; however, in most applications, it is the case that, once deployed, these algorithms will encounter data similar to, but different from the training set. The focus of our work here is on domain adaptation. That is, we are interested in developing data-driven performance bounds for scenarios where there exists some difference between training and test distributions. These bounds are critical for understanding the conditions under which a classifier trained on data from one distribution can perform well on data drawn from another distribution; for understanding how these conditions change as a function of the difference between training and test distributions; and for constructing algorithmic tools to improve performance. To that end, in this paper, we derive a bound on the test error as a function of a new probability distance between training and test distributions. Further, we show that this measure can be estimated using the non-parametric Friedman–Rafsky statistic, originally proposed for the multivariate two-sample test [17,18].

Domain adaptation has been a topic of some interest in the literature recently, with the research focusing on practical algorithms [1–5] and, more recently, on theoretical analysis [6–8,10–12,16]. In [6,7], Ben-David et al. relate the expected error on the test data to the expected error on the training data. They restrict their analysis to a specific hypothesis class of finite complexity and show that, for the selected hypothesis class, the test error is bounded by the $\mathcal{H}$-distance between the training and test distributions. This distance measure was first introduced in [9]. However, as the authors point out, this is not a valid bound over all domain subsets, but rather on subsets over which this type of hypothesis can commit errors. In [8], the authors derive new bounds for the case where a small subset of labeled data from the test distribution is available. In [10], Mansour et al. generalize the $\mathcal{H}-$distance to the regression problem. In contrast to these models, we propose a general non-parametric bound that can be estimated without assuming an underlying model for the data and

[*] Corresponding author. Tel.: +1 480 727 6455.
E-mail address: visar@asu.edu (V. Berisha).

http://dx.doi.org/10.1016/j.sigpro.2014.09.016
0165-1684/© 2014 Elsevier B.V. All rights reserved.

without restrictions on the hypothesis class. In [11], the authors present a new theoretical analysis of the multi-source domain adaptation problem based on the Renyi divergence. In particular, they develop optimal linear combination learning rules and characterize performance using the Renyi divergence. Although helpful to understand performance, the Renyi divergence is difficult to compute in general [13]; as we will show, the bound here can be estimated without ever computing density estimates.

We use the derived bound to construct a new active learning algorithm in a domain adaptation setting. In active learning, an algorithm interactively queries an oracle for the label of the data point that provides the greatest reduction in the test error. This topic has received a great deal of attention in the traditional learning literature where the assumption is that training and test data are drawn from the same distribution. The survey article in [14] provides an overview. In the context of domain adaptation, active learning has received less attention. The authors in [15] propose a pool-based active learning algorithm that uses the confidence of a pre-trained classifier to identify sets of points for labeling. In [22], the authors propose an online version of a similar algorithm and extend it by ruling out data points similar to the training set. Here we propose a multi-criteria cost function for active learning motivated by the new bound and evaluate it on 12 speech classification tasks. Results indicate that the proposed approach consistently yields a lower error rate than a competing alternative.

2. Domain adaptation bound

We consider a binary classification problem. Let us define data from two domains, the source (training) and the target (testing) domain and a corresponding labeling function for each domain $g_S, g_T : \mathbf{x} \to \{0,1\}$ that yields the true class label of a given data point. The source domain, denoted by the pair $(\mathbf{X}_S, g_S(\mathbf{X}_S))$, represents the data used to train the machine learning algorithm and the data $(\mathbf{X}_T, g_T(\mathbf{X}_T))$ represents the data the algorithm will encounter once deployed. The rows of the source and target data are drawn from $f_S(\mathbf{x})$ and $f_T(\mathbf{x})$. The risk, or the probability that a hypothesis, h, disagrees with the true labeling function is defined as

$$\epsilon_S(h,g_S) = \mathbf{E}_{f_S(x)}[|h(\mathbf{x}) - g_S(\mathbf{x})|], \tag{1}$$

for the source data. It is similarly defined for the target data. In Theorem 1, we identify a relationship between the error rates on the source and target data.

Theorem 1. *Given a hypothesis, h, the target error, $\epsilon_T(h,g_T)$, can be bounded by the error on the source data, $\epsilon_S(h,g_S)$, the difference between labeling functions, and a distance measure between source and target distributions as follows*:

$$\epsilon_T(h,g_T) \le \epsilon_S(h,g_S) + \mathbf{E}_{f_S(x)}[|g_S(\mathbf{x}) - g_T(\mathbf{x})|] + 2\mathcal{D}_{FR}(f_S,f_T), \tag{2}$$

where $\mathcal{D}_{FR}(f_S,f_T) = \sqrt{\frac{1-\int (f_S(\mathbf{x}) f_T(\mathbf{x})}{(0.5 f_S(\mathbf{x}) + 0.5 f_T(\mathbf{x}))) \, d\mathbf{x}}}.$

Proof of Theorem 1. The proof begins in the same fashion as the result in [7] and then diverges:

$$\epsilon_T(h,g_T) = \epsilon_T(h,g_T) + \epsilon_S(h,g_S) - \epsilon_S(h,g_S) + \epsilon_S(h,g_T) - \epsilon_S(h,g_T) \tag{3}$$

$$\epsilon_T(h,g_T) \le \epsilon_S(h,g_S) + |\epsilon_S(h,g_T) - \epsilon_S(h,g_S)| + |\epsilon_T(h,g_T) - \epsilon_S(h,g_T)| \tag{4}$$

$$\epsilon_T(h,g_T) \le \epsilon_S(h,g_S) + \mathbf{E}_{f_S(x)}[|g_S(\mathbf{x}) - g_T(\mathbf{x})|] + \left| \int f_T(\mathbf{x}) |h(\mathbf{x}) - g_T(\mathbf{x})| \right| d\mathbf{x} - \int f_S(\mathbf{x}) |h(\mathbf{x}) - g_T(\mathbf{x})| \, d\mathbf{x}| \tag{5}$$

$$\epsilon_T(h,g_T) \le \epsilon_S(h,g_S) + \mathbf{E}_{f_S(x)}[|g_S(\mathbf{x}) - g_T(\mathbf{x})|] + \int |f_T(\mathbf{x}) - f_S(\mathbf{x})| |h(\mathbf{x}) - g_T(\mathbf{x})| \, d\mathbf{x} \tag{6}$$

$$\epsilon_T(h,g_T) \le \epsilon_S(h,g_S) + \mathbf{E}_{f_S(x)}[|g_S(\mathbf{x}) - g_T(\mathbf{x})|] + \int | f_T(\mathbf{x}) - f_S(\mathbf{x})| \, d\mathbf{x} \tag{7}$$

Going from (6) to (7), we recognize that the maximum value of $|h(\mathbf{x}) - g_T(\mathbf{x})|$ is 1. In (7), we identify an upper bound on the target error expressed using the Kolmogorov total variation (TV) distance between source and target distributions, $\mathcal{D}_1(f_S,f_T) = \frac{1}{2}\int |f_T(\mathbf{x}) - f_S(\mathbf{x})| \, d\mathbf{x}$.
If we let

$$A(f_S,f_T) = \int \frac{f_S(\mathbf{x}) f_T(\mathbf{x})}{0.5 f_S(\mathbf{x}) + 0.5 f_T(\mathbf{x})} \, d\mathbf{x}, \tag{8}$$

then

$$1 - A(f_S,f_T) = \int 0.5 f_S(\mathbf{x}) + 0.5 f_T(\mathbf{x}) \, d(\mathbf{x}) - \int \frac{f_S(\mathbf{x}) f_T(\mathbf{x})}{0.5 f_S(\mathbf{x}) + 0.5 f_T(\mathbf{x})} \, d\mathbf{x} \tag{9}$$

$$1 - A(f_S,f_T) = \int \frac{[0.5 f_S(\mathbf{x}) - 0.5 f_T(\mathbf{x})]^2}{0.5 f_S(\mathbf{x}) + 0.5 f_T(\mathbf{x})} \, d\mathbf{x}. \tag{10}$$

An upper bound on the total variation distance can be expressed in terms of $A(f_S,f_T)$ as follows:

$$\mathcal{D}_1(f_S,f_T) = \frac{1}{2}\int |f_T(\mathbf{x}) - f_S(\mathbf{x})| \, d\mathbf{x} \tag{11}$$

$$\mathcal{D}_1(f_S,f_T) = \int \left| \frac{0.5 f_T(\mathbf{x}) - 0.5 f_S(\mathbf{x})}{\sqrt{0.5 f_S(\mathbf{x}) + 0.5 f_T(\mathbf{x})}} \times \sqrt{0.5 f_S(\mathbf{x}) + 0.5 f_T(\mathbf{x})} \right| d\mathbf{x} \tag{12}$$

$$\mathcal{D}_1(f_S,f_T) \le \sqrt{1 - A(f_S,f_T)} \tag{13}$$

Going from (12) to (13), we make use of the Cauchy–Schwarz inequality and simplify the resulting terms [19]. The inequality in (13) upper bounds the total variation distance in terms of $A(f_S,f_T)$. We refer to this upper bound as the Friedman–Rafsky (FR) distance, $\mathcal{D}_{FR}(f_S,f_T) = \sqrt{1 - A(f_S,f_T)}$ (named after the non-parametric estimator corresponding to $A(f_S,f_T)$ [17,18]). □

The bound in Theorem 1 depends on three terms: the error on the source data, the expected difference in the labeling functions across the two domains, and a measure of the distance between source and target distributions (Friedman–Rafsky distance). We expect that the selected training algorithm will seek to minimize the first term; the second term characterizes the difference between labeling functions in the source and target domains; the third term

Download English Version:

https://daneshyari.com/en/article/6959865

Download Persian Version:

https://daneshyari.com/article/6959865

Daneshyari.com