A Practical Approach to Reduce the Learning Bias Under Covariate Shift

Van-Tinh Tran(✉) and Alex Aussem

LIRIS, UMR 5205, University of Lyon 1, 69622 Lyon, France
{van-tinh.tran,aaussem}@univ-lyon1.fr

Abstract. Covariate shift is a specific class of selection bias that arises when the marginal distributions of the input features X are different in the source and the target domains while the conditional distributions of the target Y given X are the same. A common technique to deal with this problem, called importance weighting, amounts to reweighting the training instances in order to make them resemble the test distribution. However this usually comes at the expense of a reduction of the effective sample size. In this paper, we show analytically that, while the unweighted model is globally more biased than the weighted one, it may locally be less biased on low importance instances. In view of this result, we then discuss a manner to optimally combine the weighted and the unweighted models in order to improve the predictive performance in the target domain. We conduct a series of experiments on synthetic and real-world data to demonstrate the efficiency of this approach.

1 Introduction

Selection bias, also termed dataset shift or domain adaptation in the literature [8], occurs when the training distribution $P(x, y)$ and the test distribution $P'(x, y)$ are different. It is pervasive in almost all empirical studies, including Machine Learning, Statistics, Social Sciences, Economics, Bioinformatics, Biostatistics, Epidemiology, Medicine, etc. Selection bias is prevalent in many real-world machine learning problems because the common assumption in machine learning is that the training and the test data are drawn independently and identically from the same distribution. The term "domain adaptation" is used when one builds a model from some fixed source domain, but wishes to deploy it across one or more different target domains. The term "selection bias" is slightly more specific as it assumes implicitly that there exists a binary variable S that controls the selection of examples in the training set, in other words we only have access to the examples that have $S = 1$. For instance, case-control studies in Epidemiology are particularly susceptible to selection bias, including bias resulting from inappropriate selection of controls in case-control studies, bias resulting from differential loss-to-follow-up, incidence-prevalence bias, volunteer bias, healthy-worker bias, and nonresponse bias [4].

It is well known that one may account for the difference between $P(x, y)$ and $P'(x, y)$ by re-weighting the training points using the so-called importance

A. Appice et al. (Eds.): ECML PKDD 2015, Part II, LNAI 9285, pp. 71–86, 2015.
DOI: 10.1007/978-3-319-23525-7_5

weight, denoted as $\beta(x,y) = P'(x,y)/P(x,y)$. Formally, let $\{h_{\theta^*}\}_{\theta \in \Theta}$ be a model family from which we want to select an optimal model $h_{\theta^*}(x) = h(x, \theta^*)$ for our learning task and let $l(y, h(x,\theta))$ be the loss function we would like to minimize, the optimal model we are searching for is the one that minimizes the expected loss over the test (or target) distribution:

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} \sum_{(x,y)\sim P} \beta(x,y) P(x,y) l(y, h(\theta, x))$$

So in practice, weighting the empirical loss of the training instances by $\beta(x,y)$ provides a well-justified solution to the selection bias problem.

In general, the estimation of $\beta(x,y)$ with two different distributions $P(x,y)$ and $P'(x,y)$ is unsolvable, as the two terms could be arbitrarily far apart. One simple assumption we can make about the connection between the distributions of the source and the target domains is that $P(x,y)$ and $P'(x,y)$ differ only in $P(x)$ and $P'(x)$ while their conditional distribution $P(y|x)$ remains unchanged. This specific selection bias is known as *covariate shift* in the literature [10]. In this case, the weighting term reduces to $\beta(x) = P'(x)/P(x)$ and effective adaptation is possible. At first glance, it may appear that covariate shift is not a problem because, for classification, we are only interested in $P(Y|X)$ which remains unchanged. In fact, Shimodaira [10] showed that there are circumstances under which the predictive performance is jeopardized by covariate shift. This happens typically when the parametric model family $\{P(Y|X,\theta)\}_{\theta \in \Theta}$ is misspecified, that is, there does not exist any $\theta \in \Theta$ such that $P(Y|X = x, \theta) = P(Y|X = x)$ for all $x \in \mathcal{X}$, so none of the models in the model family can exactly match the true relation between X and Y.

The intuitive reason why covariate shift under model misspecification is a problem is that the optimal (misspecified) model performs better in dense regions of the input space than in sparse regions, because the dense regions dominate the average classification error, which is what we want to minimize. If the dense regions of X are different in the training and test sets, the optimal model on the training set will no longer be optimal on the test set. In other words, the optimal model depends on $P(x)$, and if $P'(x) \neq P(x)$, then the optimal model for the target domain differs from that for the source domain. It was proven that, if the support of $P'(x)$ (the set of x for which $P'(x) > 0$) is contained in the support of $P(x)$, then the optimal model that maximizes this re-weighted log likelihood function asymptotically converges to the optimal model for the target domain [10] and a large body of research has been devoted to the estimation of $P'(x)/P(x)$ e.g. [13], [5], [11], [2], [1], [6], [7], [9]. However, reweighting methods do not necessarily improve the prediction accuracy as they also dependent on the extent to which the model is misspecified [12].

In this paper, we show analytically that, despite the fact that the unweighted model is globally more biased than the weighted one, the former may locally be less biased on low importance instances. In view of this result, we design a simple algorithm that combines the weighted and the unweighted models in order to improve the predictive performance in the target domain. More specifically, we prove that an optimal $B^\star$ always exists such that, in the region where $\beta(x) \leq B^\star$, the biased model trained on the unweighted sample should be preferred to the unbiased one, and vice-versa. We propose a practical procedure to estimate this threshold value from training data.

The remainder of this paper is structured as follows. In Section 2, we define some key concepts used along the paper and state some results that will support our analysis. Then in Section 3, we conduct a theoretical analysis to prove that an optimal (but not necessarily unique) $B^\star$ always exists and discuss a manner to optimally combine the weighted and the unweighted models in order to improve the predictive performance in the target domain. In section 4, a series of experiments are carried out on toy problems and real-world data sets to assess the effectiveness of this approach.

2 Preliminaries

In this section, we define some key concepts used along the paper and state some results that will support our analysis. Consider the supervised learning problem where we observed n training samples, denoted by $((x_t; y_t) : t = 1, ..., n)$, where $x_t \in \mathcal{X} \subset \mathcal{R}^d$ are i.i.d training input points drawn from some probability distribution $p(x)$ and $y_t \in \mathcal{Y} \subset \mathcal{R}$ are the corresponding training output values drawn from a conditional probability distribution $p(y|x)$. We are interested in predicting the output value y at an input point x using a model $h_\theta(x) = h(x, \theta)$ parameterized by $\theta \in \Theta \subset \mathcal{R}^m$. Under covariate shift assumption, the test inputs follow a different probability distribution $p'(x)$ while the conditional probability distribution of test output $p(y|x)$ remains unchanged. The ratio $\beta(x) = \frac{p'(x)}{p(x)}$ is called the *importance* of x. Given a loss function $l(y, h(x, \theta)) : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty)$, we shall consider throughout this paper, the following loss functions:

- **EL-Tr**: Expectation of loss over training distribution $p(x, y) = p(x)p(y|x)$

$$Loss_0(h_\theta) = E_{x,y\sim p}[l(y, h(x, \theta))] = \int p(x) \int p(y|x) l(y, h(x, \theta)) dy dx$$

- **EL-Te**: Expectation of loss over test distribution $p'(x, y) = p'(x)p(y|x)$

$$Loss_1(h_\theta) = E_{x,y\sim p'}[l(y, h(x, \theta))] = \int p'(x) \int p(y|x) l(y, h(x, \theta)) dy dx$$

- **EL-IWTr**: Expectation of Importance-weighted loss over training distribution

$$Loss_\beta(h_\theta) = E_{x,y\sim p}[\beta(x) l(y, h(x, \theta))]$$

- **B-LEL-Te**: We then define Local Expectation of loss over test distribution given $\beta(x) \leq B$ of any given hypothesis h_θ:

$$loss(h_\theta, \beta(x) \leq B) = \int_{\beta(x)\leq B} p'(x) \int_{\mathcal{Y}} p(y|x) l(y, h(x, \theta)) dy dx$$

We also define the optimal parameters of EL-Tr, EL-Te and EL-IWTr:

$$\begin{cases} \theta_0 & = \operatorname{argmin}_\theta Loss_0(h_\theta) \\ \theta_1 & = \operatorname{argmin}_\theta Loss_1(h_\theta) \\ \theta_\beta & = \operatorname{argmin}_\theta Loss_\beta(h_\theta). \end{cases}$$

It may easily be shown that EL-IWTr is equal to EL-Te,

$$\begin{aligned} E_{x,y \sim p}[\beta(x) l(y, h(x,\theta))] &= \int p(x) \int p(y|x) \frac{p'(x)}{p(x)} l(y, h(x,\theta)) dy dx \\ &= \int p'(x) \int p(y|x) l(y, h(x,\theta)) dy dx \end{aligned}$$

Therefore, minimizing EL-IWTr is equivalent to minimizing EL-Te. Nonetheless, while h_{θ_β} is globally less biased than h_{θ_0}, we will show next that it is more biased than h_{θ_0} on low-importance instances. Note that B-LEL-Te can be rewritten as:

$$loss(h_\theta, \beta(x) \leq B) = \int_{\beta(x) \leq B} \beta(x) \int_{\mathcal{Y}} p(x) p(y|x) l(y, h(x,\theta)) dy dx$$

Suppose $\beta(x)$ takes on continuous value in $[b_0, b_M]$ where $b_0 > 0$, we may rewrite B-LEL-Te as following:

$$loss(h_\theta, \beta(x) \leq B) = \int_{b_0}^{B} b \int_{\beta(x)=b} \int_{\mathcal{Y}} p(x) p(y|x) l(y, h(x,\theta)) dy dx db$$

Let $\mathcal{L}(h_\theta, \beta(x) = b) = \int_{\beta(x)=b} \int_{\mathcal{Y}} p(x) p(y|x) l(y, h(x,\theta)) dy dx$, then:

$$loss(h_\theta, \beta(x) \leq B) = \int_{b_0}^{B} b \mathcal{L}(h_\theta, \beta(x) = b) db$$

Similarly, if $\beta(x)$ takes on discrete values in $\{b_i\}_{i=0}^{M}$ such that $b_0 < b_1 < ... < b_M$, we rewrite B-LEL-IWTr as:

$$loss(h_\theta, \beta(x) \leq B) = \sum_{i=0}^{k(B)} b_i \mathcal{L}(h_\theta, \beta(x) = b_i)$$

where $k(B)$ is the largest integer such that $b_{k(B)} \leq B$. From the definitions above, we may write

$$\begin{cases} Loss_1(h_\theta) & = loss(h_\theta, \beta(x) \leq b_M), \\ Loss_0(h_\theta) & = \int_{b_0}^{\infty} \mathcal{L}(h_\theta, \beta(x) = b) db, \text{ for continuous } \beta(x), \\ Loss_0(h_\theta) & = \sum_{i=0}^{M} \mathcal{L}(h_\theta, \beta(x) = b_i), \text{ for discrete } \beta(x). \end{cases}$$

As aforementioned, a model $h(x, \theta)$ is said to be *correctly specified* if there exist parameter $\theta^* \in \Theta$ such that $h(x, \theta^*) = f(x)$, otherwise it is said to be *misspecified*. It is obvious that if a model is correctly specified, the optimal parameter θ of EL-Tr, EL-Te, and any B-LEL-Te coincide. Therefore, the model that minimizes EL-Tr will perform well on the test data globally (i.e., minimizing EL-Te) as well as locally (i.e., B-LEL-Te) in any region of the form $\beta(x) < B$. Yet, in practice, almost all models are more or less misspecified. So minimizing EL-Tr θ_0 is not necessarily equivalent minimizing EL-Te. Since EL-Te is equal to EL-IWTr, the parameter minimizing of EL-IWTr θ_β, which can be estimated from data, will also minimize EL-Te as shown in [10], [13]. However, due to the model misspecification, θ_β does not necessarily minimize B-LEL-Te. In fact, we will prove that there exist some $B^*(h_{\theta_\beta}) \in [b_0, b_M]$ such that B-LEL-Te of θ_β exceeds that of θ_0 by proving a stronger conclusion that for all model h_θ, with $\theta \in \Theta$, there exist some $B^*(h_\theta) \in [b_0, b_M]$ such that B*-LEL-Te of h_θ exceeds that of h_{θ_0}, in other words any h_θ is **locally more biased** than h_{θ_0} when predicting instance with $\beta(x) \leq B^*$.

In addition, the estimation of θ_β may subject to high variance since it involves instance weighting, which is known to reduce the effective samples size [2], [3]. Hence the idea to use h_{θ_0} of instead of h_{θ_β} to predict the test instances with $\beta(x) \leq B^*$.

3 Problem Analysis

In this section, we conduct theoretical analyses for a simple and then a more general selection bias mechanism. Those analyses will be used to derive a practical procedure aiming at reducing the bias due to covariate shift with misspecified regression or classification learning models.

We first show how EL-Tr is related to B-LEL-Te,

Lemma 1. *Suppose $\beta(x)$ takes on continuous value in $[b_0, b_M]$ with $b_M > b_0 > 0$, then:*

$$Loss_0(h_\theta) = \frac{1}{b_M} loss(h_\theta, \beta(x) \leq b_M) + \int_{b_0}^{b_M} \frac{1}{B^2} loss(h_\theta, \beta(x) \leq B) dB$$

Proof. For continuous $\beta(x)$:

$$\int_{b_0}^{b_M} \frac{1}{B^2} loss(h_\theta, \beta(x) \leq B) dB = \int_{b_0}^{b_M} loss(h_\theta, \beta(x) \leq B) d\left(\frac{-1}{B}\right)$$
$$= loss(h_\theta, \beta(x) \leq B) \left(\frac{-1}{B}\right) |_{b_0}^{b_M} - \int_{b_0}^{b_M} \frac{-1}{B} d(loss(h_\theta, \beta(x) \leq B))$$

By definition, $loss(h_\theta, \beta(x) \leq B) = \int_{b_0}^{B} b\mathcal{L}(b, h_\theta)db$, so $loss(h_\theta, \beta(x) \leq b_0) = 0$ and $d(loss(h_\theta, \beta(x) \leq B)) = B\mathcal{L}(h_\theta, \beta(x) = B)dB$. Thus:

$$\begin{aligned}\int_{b_0}^{b_M} \frac{1}{B^2} loss(h_\theta, \beta(x) \leq B)dB &= \frac{-1}{b_M} loss(h_\theta, \beta(x) \leq b_M) \\ &\quad + \int_{b_0}^{b_M} \frac{1}{B}(B\mathcal{L}(h_\theta, \beta(x) = B)dB)\end{aligned}$$

By definition, we have $Loss_0(h_\theta) = \int_{b_0}^{b_M} \mathcal{L}(h_\theta, \beta(x) = B)dB$, so:

$$\int_{b_0}^{b_M} \frac{1}{B^2} loss(h_\theta, \beta(x) \leq B)dB = -\frac{1}{b_M} loss(h_\theta, b_M) + Loss_0(h_\theta)$$

which conclues the proof □

A similar results holds in the discrete case.

Corollary 1. *Suppose $\beta(x)$ takes on discrete values $\{b_i\}_{i=0}^{M}$ such that $b_0 < b_1 < ... < b_M$, then:*

$$Loss_0(h_\theta) = \frac{1}{b_M} loss(h_\theta, \beta(x) \leq b_M) + \sum_{k=0}^{M-1} \left(\frac{1}{b_k} - \frac{1}{b_{k+1}}\right) loss(h_\theta, \beta(x) \leq b_k)$$

Proof.

$$\begin{aligned}&\sum_{k=0}^{M-1} \left(\frac{1}{b_k} - \frac{1}{b_k + 1}\right) loss(h_\theta, \beta(x) \leq b_k) + \frac{1}{b_M} loss(h_\theta, \beta(x) \leq b_M) \\ &= \left(\frac{1}{b_0} - \frac{1}{b_1}\right)[b_0\mathcal{L}(h_\theta, \beta(x) = b_0)] \\ &\quad + \left(\frac{1}{b_1} - \frac{1}{b_2}\right)[b_0\mathcal{L}(h_\theta, \beta(x) = b_0) + b_1\mathcal{L}(h_\theta, \beta(x) = b_1)] \\ &\quad + ... \\ &\quad + \left(\frac{1}{b_{M-1}} - \frac{1}{b_M}\right)[b_0\mathcal{L}(h_\theta, \beta(x) = b_0) + ... + b_{M-1}\mathcal{L}(h_\theta, \beta(x) = b_{M-1})] \\ &\quad + \frac{1}{b_M}[b_0\mathcal{L}(h_\theta, \beta(x) = b_0) + b_1\mathcal{L}(h_\theta, \beta(x) = b_1) + .. + b_M\mathcal{L}(h_\theta, \beta(x) = b_M)]\end{aligned}$$

$$
\begin{aligned}
&= b_0\mathcal{L}(h_\theta, \beta(x) = b_0)\left[\left(\frac{1}{b_0} - \frac{1}{b_1}\right) + \left(\frac{1}{b_1} - \frac{1}{b_2}\right) + ... + \left(\frac{1}{b_{M-1}} - \frac{1}{b_M}\right) + \frac{1}{b_M}\right] \\
&\quad + ... \\
&\quad + b_{M-1}\mathcal{L}(h_\theta, \beta(x) = b_{M-1})\left[\left(\frac{1}{b_{M-1}} - \frac{1}{b_M}\right) + \frac{1}{b_M}\right] \\
&\quad + b_M\mathcal{L}(h_\theta, \beta(x) = b_M)\left[\frac{1}{b_M}\right] \\
&= \sum_{i=0}^{M} \mathcal{L}(h_\theta, \beta(x) = b_i) = Loss_0(h_\theta) \qquad \square
\end{aligned}
$$

In view of Corollary 1, we may now state the following theorem,

Theorem 1. *Suppose there exists two real values, b_0 and b_1, such that $b_0 < 1 < b_1$ and a subset $X_0 \subset \mathcal{X}$ such that*

$$
\beta(x) = \begin{cases} b_0 & \textit{if } x \in X_0 \\ b_1 & \textit{if } x \notin X_0, \end{cases}
$$

then there exists a threshold B^* such that:

$$
loss(h_{\theta_1}, \beta(x) \leq B^*) \geq loss_1(h_{\theta_0}, \beta(x) \leq B^*).
$$

In fact, B^* can take any value in $[b_0, b_1)$.

Proof. By definition, $Loss_0(h_{\theta_0}) \leq Loss_0(h_{\theta_1})$, using Lemma 1, we may write:

$$
\begin{aligned}
Loss_0(h_{\theta_0}) &= \frac{1}{b_1} loss(h_{\theta_0}, \beta(x) \leq b_1) + \left(\frac{1}{b_0} - \frac{1}{b_1}\right) loss(h_{\theta_0}, \beta(x) \leq b_0) \\
&= \frac{1}{b_1} Loss_1(h_{\theta_0}) + \left(\frac{1}{b_0} - \frac{1}{b_1}\right) loss(h_{\theta_0}, \beta(x) \leq b_0)
\end{aligned}
$$

Similarly,

$$
Loss_0(h_{\theta_1}) = \frac{1}{b_1} Loss_1(h_{\theta_1}) + \left(\frac{1}{b_0} - \frac{1}{b_1}\right) loss(h_{\theta_1}, \beta(x) \leq b_0)
$$

Thus,

$$
\begin{aligned}
\frac{1}{b_1} Loss_1(h_{\theta_0}) + \left(\frac{1}{b_0} - \frac{1}{b_1}\right) loss(h_{\theta_0}, \beta(x) \leq b_0) &\leq \frac{1}{b_1} Loss_1(h_{\theta_1}) \\
+ \left(\frac{1}{b_0} - \frac{1}{b_1}\right) loss(h_{\theta_1}, \beta(x) \leq b_0)
\end{aligned}
$$

Finally,

$$
loss(h_{\theta_1}, \beta(x) \leq b_0) - loss(h_{\theta_0}, \beta(x) \leq b_0) = \frac{b_0}{b_1 - b_0}\left[Loss_1(h_{\theta_0}) - Loss_1(h_{\theta_1})\right]
$$

It is easily shown that the right hand side of inequality above is non-negative due to the definition of θ_1. It follows that

$$loss(h_{\theta_1}, \beta(x) \leq b_0) - loss(h_{\theta_0}, \beta(x) \leq b_0) \leq 0$$

which, given the assumption about $\beta(x)$, is equivalent to,

$$loss(h_{\theta_1}, \beta(x) = b_0) - loss(h_{\theta_0}, \beta(x) = b_0) \leq 0$$

Thus the Theorem is true when $B^* = b_0$. It is also true for any other $B^* \in [b_0, b_1)$ as a consequence. □

When the assumptions of Theorem 1 holds, we say that the covariate shift scheme follows a simple step distribution. The equality in Theorem 1 only occurs when θ_0 minimizes EL-Te and θ_1 minimizes EL-Tr. Such condition indicates that covariate shift does not have an effect on searching for optimal θ, which is a rare case as shown by other studies. Theorem 1 shows that for *simple step distribution* where inclusion in the training sample is either proportional to b_0^{-1} (over-sampled instances), or to b_1^{-1} (under-sampled instances), h_{θ_0} exhibits a lower bias compared to h_{θ_1} on the low importance test instances. This type of selection bias mechanism is actually quite common. For instance, prospective cohort studies in epidemiology are by design prone to covariate shift because selection criteria are associated with the exposure to potential risk factors.

Theorem 2. *For all $\theta \in \Theta$, there exists a threshold $B^*(h_\theta)$ such that*

$$loss(h_\theta, \beta(x) \leq B^*(h_\theta)) \geq loss(h_{\theta_0}, \beta(x) \leq B^*(h_\theta)) \tag{1}$$

$B^*(h_\theta)$ could take any value in the set below:

$$\mathcal{B}^*(h_\theta) = \underset{B}{\operatorname{argmax}}(loss(h_\theta, \beta(x) \leq B) - loss(h_{\theta_0}, \beta(x) \leq B))$$

The equality occurs whenever θ_1 is also a minimum for EL-Tr.

Proof. We prove by contradiction that Theorem 2 holds. Assume that inequality 1 does not hold for $B^*(h_\theta)$ defined above:

$$loss(h_\theta, \beta(x) \leq B^*(h_\theta)) - loss(h_{\theta_0}, \beta(x) \leq B^*(h_\theta)) < 0 \tag{2}$$

By definition of $B^*(h_\theta)$, we may show that, for all $B \in [b_0, b_M]$,

$$loss(h_\theta, \beta(x) \leq B) - loss(h_{\theta_0}, \beta(x) \leq B) < 0$$

Thus, for all $B \in [b_0, b_M]$

$$loss(h_{\theta_0}, \beta(x) \leq B) > loss(h_\theta, \beta(x) \leq B)$$

Now, using Lemma 1 for continuous $\beta(x)$, we have:

$$\begin{aligned} Loss_0(h_{\theta_0}) &= \frac{1}{b_M} loss(h_{\theta_0}, \beta(x) \leq b_M) + \int_{b_0}^{b_M} \frac{1}{B^2} loss(h_{\theta_0}, \beta(x) \leq B) dB \\ &> \frac{1}{b_M} loss(h_\theta, \beta(x) \leq b_M) + \int_{b_0}^{b_M} \frac{1}{B^2} loss(h_\theta, \beta(x) \leq B) dB = Loss_0(h_\theta) \end{aligned}$$

Hence, $Loss_0(h_{\theta_0}) > Loss_0(h_\theta)$, contradicts the fact that $\theta_0 = \text{argmin}_\theta$ $Loss_0(h_\theta)$ is the optimal hypothesis under the unweighting scheme and $\theta \neq$ $\text{argmin}_\theta\, Loss_0(h_\theta)$.

If the two terms in inequality 1 are equal, then we can prove similarly that $Loss_0(h_{\theta_0}) = Loss_0(h_\theta)$, which implies that θ_1 is also a minimal solution of EL-Tr. The demonstration for discrete $\beta(x)$ values follows similarly. □

Theorem 2 states that any model h_θ with $\theta \in \Theta$ is outperformed by h_{θ_0} learned from the unweighted training samples in terms of bias when predicting examples with $\beta(x) \leq B^*(h_\theta)$. This is also applied to model h_{θ_β} which minimizes EL-IWTr. In addition, the estimation of θ_β may exhibit a higher variance due to the effective sample size reduction as discussed in [2,3]. These results altogether suggest that h_{θ_0} should be preferred to h_{θ_β} for predicting the instance's outputs in the region $\beta(x) \leq B^*(h_\theta)$, termed **low-importance region**. Therefore, for any learning task with covariate shift, we shall train two distinct models, one with and the other without the importance weighting scheme. Then, we shall use the latter to predict instances satisfying $\beta(x) \leq B^*(h_\theta)$ and use the former to predict the remaining instances. The optimal value for $B^*(h_\theta)$ may be estimated from the training data. The set of all possible empirical threshold $\hat{B}^*(h_{\theta_\beta})$ can be obtained empirically by solving the following problem :

$$\hat{\mathcal{B}}^*(h_\theta) = \underset{B}{\text{argmax}}\, \frac{1}{n} \sum_{\substack{i \in \{1,..,n\} \\ \beta(x_i) \leq B}} \beta(x_i)[l(y_i, h(x_i, \theta_\beta)) - l(y_i, h(y_i, \theta_0))] \qquad (3)$$

As n grows to infinity, it follows from the law of large numbers that,

$$\hat{\mathcal{B}}^*(h_\theta) \to \mathcal{B}^*(h_\theta)$$

Therefore, $B^*(h_{\theta_\beta})$ could be estimated empirically either from training data or by cross validation. In this study, we use a 5-fold importance weighted cross validation to estimate $B^*(h_{\theta_\beta})$ as suggested in [11]. It should be emphasized that $B^*(h_{\theta_\beta})$ is not necessarily unique. For instance, any value between b_0 and b_1 in Theorem 1 is admissible as mentioned earlier.

4 Experiments

In this section, we assess the ability of our "hybrid approach" to reduce the learning bias under covariate shift based on Theorem 2. We first discuss the strategies employed to estimate the importance weights: one is based explicitly on the true bias mechanism, the other is based on linear density-ratio model. We emphasize that the latter does not require any prior knowledge of the true sampling probabilities to estimate the $\beta(x)$ values, and uses the test input features instead. In fact, the estimation of distribution is a hard problem, thus it is more appealing to directly estimate $\beta(x)$. Indeed, a large body of work has been devoted to this line of research e.g. [13], [5], [11], [9], [2], [1], [6]. From the many references, we choose the Unconstrained Least-Square Importance Fitting (uLSIF) estimator for $\beta(x)$ that was proved to be successful with covariate shift. We then study a toy regression problem to show if covariate shift corrections based on our method reduces prediction error on the test set when the learning model is misspecified. We then test our approach on real world benchmark data sets, from which the training examples are selected according to various biased sampling schemes as suggested in [6].

4.1 Importance Ratio Estimation

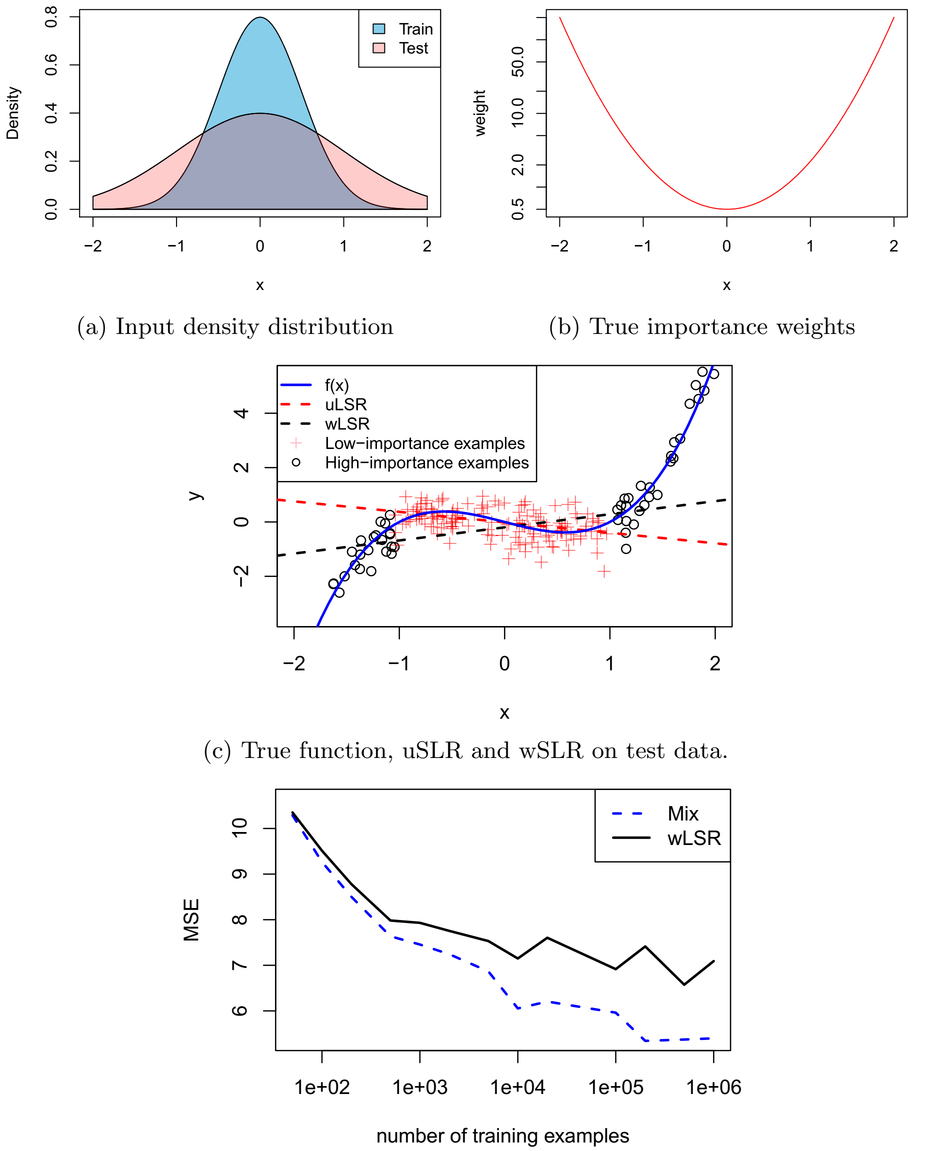
As aforementioned, we use two weighting schemes in ours experiments, one is derived from the true selection bias mechanism an one is Unconstrained Least-Square Importance Fitting (uLSIF), a method based on linear density-ratio models [6]. Formally, it assumes that the density ratio $\beta(x)$ can be approximated by a linear model $\hat{\beta}(x) = \sum_{i=1}^{M} \alpha_i h_i(x)$ where the basis functions h_i, $i = 1, ..., M$ are chosen so that $h_i(x) \geq 0$ for all input value x. The coefficients $\alpha_1, ..., \alpha_M$ are parameters of the linear model and are estimated from data by minimizing the empirical square error between weighted biased distribution (from training data) and the bias-free distribution of x:

$$\min_{\alpha} \frac{1}{2n} \sum_{i=1}^{n} (\hat{\beta}(x_i))^2 - \frac{1}{n'} \sum_{i=1}^{n'} \hat{\beta}({x'}_i) + \lambda.Reg(\alpha)$$

where $\{x_i\}_{i=1}^{n}$ and $\{x'_i\}_{i=1}^{n'}$, are the training and test inputs, $Reg(\alpha)$ is the regularization term, introduced to avoid overfitting. A heuristic choice of $h_i(x)$ proposed in [6] is a Gaussian kernel centered at the test points $\{x_i\}_{i=1}^{n'}$ when the number of test points is small (less than 100) or at *template* points $\{x'_i\}_{i=1}^{100}$, which is a random subset of test set when the number of test points is large for computation advantage. The kernel width and the regularization term $Reg(\alpha)$ are optimized by cross-validation with grid search.

4.2 Toy Regression Problem

Consider the following training data generating process: $x \sim N(\mu_0, \sigma_0)$ and $y = f(x) + \epsilon$, where $\mu_0 = 0$, $\sigma_0 = 0.5$, $f(x) = -x + x^3$, and $\epsilon \sim N(0, 0.3)$. In the test data, we have the same relationship between x and y but the distribution of the covariate x is shifted to $x \sim N(\mu_1, \sigma_1)$, where $\mu_1 = 0$, $\sigma_1 = 1$. The training and test distributions, along with their ratio are depicted in Fig. 1a and 1b. The minimization of EL-Tr is obtained using the unweighted Least Square Regression (uLSR) method for the normal regression while minimization of EL-Te is performed by the weighted Least Square Regression (wLSR). As shown in [10], wLSR is unbiased thus it should perform better than uLSR, which is biased, on test data. However, as can be seen in Fig. 1c, uLSR (red dashed line) seems to better approximate the $y = f(x)$ curve (in blue) than wLSR (black dashed line) on instances in the interval $(-1, 1)$. As may be seen in Fig.1d, the hybrid model that optimally combines wLSR and uLSR, based



(a) Input density distribution

(b) True importance weights

(c) True function, uSLR and wSLR on test data.

(d) MSE vs training sample size

Fig. 1. An illustrative example of fitting a function f(x) using a linear model with/without the weight importance scheme (wLSR/uLSR) and a combination of both (termed "Mix").

on Theorem 1, achieves a lower Mean Square Error (MSE) compared to wLSR. The experiment was repeated 30 times for each number of sample size. It should be noted that the hybrid model always outperforms the weighted model and the gain in performance on the test set is more noticeable for larger training sizes.

4.3 Simple Step Sample Selection Distribution

In this second experiment, we consider a simple step distribution with known or estimated selection probabilities and we apply this selection scheme on a variety of UCI data sets in order to assess the efficiency of our bias correction procedure in more realistic scenarios. We use a SVM classifier for both classification and regression tasks. Experiments are repeated 100 times for each data set. In each trial, we randomly select an input feature x^c to control the bias along with 300 training samples. We then apply the following single step probability distribution as discussed in Theorem 1,

$$P(s=1|x=x_i^c) = p_s = \begin{cases} p1 = 0.9 \text{ if } x_i^c \leq mean(x^c) \\ p2 = \frac{0.9}{1+exp(r)} \text{otherwise} \end{cases}$$

where r is a parameter that controls the strength of the selection bias. In each trial r takes a random value from a normal distribution $N(2, 0.1)$. With these parameters, the selection probability for instances having an x^c value (e.g. a degree of exposure to some risk factor) above the mean is between 7 to 10 times smaller than for those having of a lower value. This is a scenario that typically arises in epidemiological cohort studies when subjects are includes in the study

Table 1. Mean test error averaged over 200 trials with different weighting schemes on 15 UCI data sets. Data sets marked with '*' are regression problems. P denotes the weighting scheme using the true selection probability and $\hat{P}$ denotes the weighting scheme using a noisy selection probability. For each pair of weighted and mix models, the better prediction value is highlighted in boldface

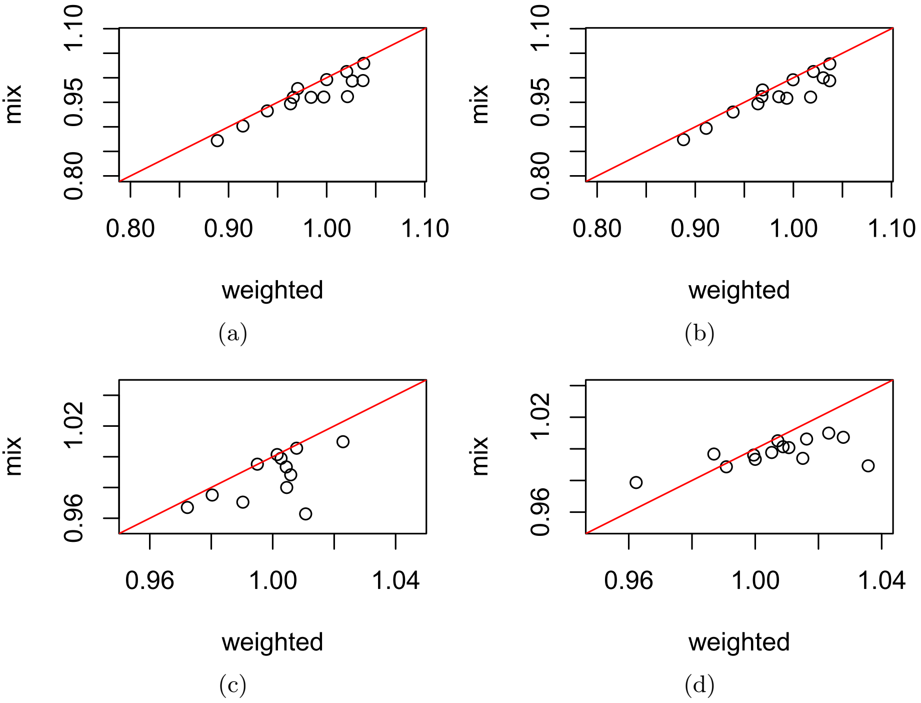
Data set	No weighting	P	P mix	$\hat{P}$	$\hat{P}$ mix
India diabetes	1.000 ± 0.020	0.966 ± 0.019	**0.960** ± 0.018	0.968 ± 0.019	**0.962** ± 0.018
Ionosphere	1.000 ± 0.128	0.915 ± 0.105	**0.902** ± 0.107	0.911 ± 0.104	**0.897** ± 0.106
BreastCancer	1.000 ± 0.039	1.020 ± 0.044	**1.013** ± 0.044	1.020 ± 0.044	**1.013** ± 0.043
GermanCredit	1.000 ± 0.008	1.000 ± 0.007	**0.996** ± 0.008	1.000 ± 0.008	**0.996** ± 0.008
Australian credit	1.000 ± 0.006	0.963 ± 0.008	**0.947** ± 0.010	0.964 ± 0.008	**0.947** ± 0.010
Mushroom	1.000 ± 0.068	0.090 ± 0.057	**0.872** ± 0.060	0.888 ± 0.058	**0.874** ± 0.056
Congressional Voting	1.000 ± 0.033	1.026 ± 0.039	**0.993** ± 0.038	1.030 ± 0.038	**1.000** ± 0.037
Banknote	1.000 ± 0.040	**0.970** ± 0.043	0.978 ± 0.038	**0.969** ± 0.042	0.975 ± 0.039
Airfoil self noise*	1.000 ± 0.023	0.997 ± 0.015	**0.961** ± 0.012	0.993 ± 0.015	**0.958** ± 0.012
Abanlone*	1.000 ± 0.032	0.984 ± 0.020	**0.960** ± 0.020	0.985 ± 0.021	**0.961** ± 0.020
Auto MGP*	1.000 ± 0.084	0.939 ± 0.066	**0.933** ± 0.067	0.939 ± 0.066	**0.930** ± 0.067
Boston Housing*	1.000 ± 0.057	1.037 ± 0.053	**0.994** ± 0.050	1.037 ± 0.053	**0.994** ± 0.050
Space GA*	1.000 ± 0.009	1.021 ± 0.007	**0.962** ± 0.008	1.018 ± 0.008	**0.961** ± 0.008
Cadata*	1.000 ± 0.013	1.038 ± 0.022	**1.029** ± 0.017	1.037 ± 0.022	**1.029** ± 0.017

according to some exposure factor. Consider the two following weighting schemes. The first one: $\beta = p'(x)/p(x) = p(s=1)/p(s=1|x) \sim 1/p_s$ assumes that the bias mechanism is known exactly.

$$\beta(x) \sim p_s^{-1} \sim \begin{cases} b1 = 1 \text{ if } x_i^c \leq mean(x^c) \\ b2 = 1 + exp(r)\text{otherwise} \end{cases}$$

In practice, however, the selection probability is rarely known exactly. So let us assume that the estimation of β is subject to some error and let us consider the following approximate weighting scheme:

$$\hat{\beta}(x) \sim p_s^{-1} \sim \begin{cases} b1 = 1 \text{ if } x_i^c \leq mean(x^c) \\ b2 = 1 + exp(\hat{r}) \text{ if otherwise} \end{cases}$$



(a)

(b)

(c)

(d)

Fig. 2. MSE gain of the mix model vs. MSE gain of the weighted model. Points below the diagonal line indicate that the mix model outperforms the weighted model. Figures (a) and (b): simple step distribution covariate shift used in the first experiment with the weighted model based on (a) the true selection probability and (b) based on the estimated selection probability. Figures (c) and (d): covariate shift in used in the second experiment when the weighted model based was based on: (c) true selection probability; (d) on uLSIF.

where $\hat{r} = r + \mathcal{N}(0, 0.1)$ is our noisy estimate of r. For each weighting scheme, we fit a true weighted model (denoted as P in Table 1) and an approximated weighted model (denoted as $\hat{P}$ in Table 1). As $p_1 < 1$ and $p_2 > 1$, our weighting mechanism satisfies the assumptions of Theorem 1, so we set $B^* = 1$. We report the mean square errors (MSE) in Table 1. All values are normalized by the MSE of the unweighted model (our gold standard). As may be seen from the plots in Fig. 2a and 2b, the combined models outperform the weighted ones. That is, when using either exact probability ratio, the results obtained with P_{mix} are better than that of P. The same observation can be made when the estimated probability ratios are used instead (i.e., $\hat{P}_{mix}$ versus $\hat{P}$) and except on the Banknote data set. The gain is significant at the significance level 5% using the Wilcoxon signed rank test.

4.4 General Covariate Selection Mechanisms

In this last experiment, we use the same setting as above but we use a more general distribution:

$$P(s = 1|x = x_i^c) = ps = \begin{cases} p1 = 0.9 \text{ if } x_i^c \leq mean(x^c) \\ p2 = 0.1 \text{ if } x_i^c > mean(x^c) + 0.8 \times 2\sigma(x^c) \\ p3 = 0.9 - \frac{x_i^c - mean(x^c)}{2\sigma(x^c)} \text{ otherwise.} \end{cases}$$

where $\sigma(x^c)$ denotes the standard deviation of x^c. As may be observed, the assumptions required in Theorem 1 do no hold anymore with this more general

Table 2. Mean test error averaged over 200 trials for different weighting schemes on UCI data set. Data sets marked with * are for regression problems. P denotes the weighting scheme based on the true selection probability and uLSIF denotes the weighting scheme using the uLSIF estimator. For each pair of weighted and mix models, the better prediction value is highlighted in boldface.

Data set	No weighting	P	P mix	uLSIF	uLSIF mix
India diabetes	1.000 ± 0.021	0.980 ± 0.018	**0.975** ± 0.018	1.016 ± 0.021	**1.006** ± 0.021
Ionosphere	1.000 ± 0.087	1.006 ± 0.087	**0.988** ± 0.085	1.028 ± 0.093	**1.007** ± 0.087
BreastCancer	1.000 ± 0.019	1.004 ± 0.018	**0.993** ± 0.019	1.000 ± 0.018	**0.993** ± 0.019
GermanCredit	1.000 ± 0.008	1.003 ± 0.008	**0.999** ± 0.008	1.009 ± 0.008	**1.001** ± 0.008
Australian credit	1.000 ± 0.009	0.972 ± 0.007	**0.967** ± 0.007	1.007 ± 0.008	**1.005** ± 0.008
Mushroom	1.000 ± 0.558	1.011 ± 0.054	**0.963** ± 0.051	0.991 ± 0.054	**0.989** ± 0.054
Congressional Voting	1.000 ± 0.037	1.023 ± 0.036	**1.010** ± 0.037	**0.987** ± 0.036	0.997 ± 0.036
Banknote	1.000 ± 0.060	1.083 ± 0.057	**0.962** ± 0.062	**0.962** ± 0.061	0.979 ± 0.058
Airfoil self noise*	1.000 ± 0.007	0.995 ± 0.007	**0.995** ± 0.007	1.011 ± 0.008	**1.001** ± 0.008
Abanlone*	1.000 ± 0.007	1.001 ± 0.008	1.001 ± 0.007	1.005 ± 0.007	**0.998** ± 0.006
Auto MGP*	1.000 ± 0.026	0.990 ± 0.025	**0.970** ± 0.025	1.015 ± 0.027	**0.994** ± 0.026
Boston Housing*	1.000 ± 0.043	0.984 ± 0.031	**0.940** ± 0.032	1.036 ± 0.040	**0.989** ± 0.042
Space GA*	1.000 ± 0.006	1.005 ± 0.005	**0.980** ± 0.006	1.000 ± 0.005	**0.996** ± 0.005
Cadata*	1.000 ± 0.012	1.008 ± 0.013	**1.006** ± 0.012	1.023 ± 0.013	**1.010** ± 0.012

sample selection distribution. According to Eq. 3, we need to estimate $\hat{B}^*(h_\theta)$ empirically from data. We consider again two importance weighting schemes: one is based on the true underlying probability and is referred to as P, while the other is based on the uLSIF estimator. As may be observed from Table 2 and Figures 2c and 2d that performances of the hybrid models are significantly improved with respect to the weighted models, except with the Congressional Voting and Banknote data sets.

5 Conclusions

In this paper, we showed that the standard importance weighting approach used to reduce the bias due to covariate shift can easily be improved when misspecified training models are used. Considering a simple class of selection bias mechanisms, we proved analytically that the unweighted model exhibits a lower prediction bias compared to the globally unbiased model in the low importance input subspace. Even for more general covariate shift scenarios, we proved that there always exist a threshold for the importance weight below which the test instances should be predicted by the globally biased model. In view of this result, we proposed a practical procedure to estimate this threshold and we discussed a simple procedure to combine the weighted and unweighted prediction models. The method was shown to be effective in reducing the bias on several UCI data sets.

Acknowledgments. This work was partially supported by a grant from the European ENIAC Joint Undertaking (INTEGRATE project).

References

1. Bickel, S., Brückner, M., Scheffer, T.: Discriminative learning under covariate shift. The Journal of Machine Learning Research **10**, 2137–2155 (2009)
2. Cortes, C., Mansour, Y., Mohri, M.: Learning bounds for importance weighting. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 23, pp. 442–450. Curran Associates, Inc. (2010)
3. Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B.: Covariate shift by kernel mean matching (2009)
4. Hernán, M.A., Hernández-Díaz, S., Robins, J.M.: A structural approach to selection bias. Epidemiology **15**(5), 615–625 (2004)
5. Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., Schölkopf, B.: Correcting sample selection bias by unlabeled data. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) NIPS, pp. 601–608. MIT Press (2006)
6. Kanamori, T., Hido, S., Sugiyama, M.: A least-squares approach to direct importance estimation. J. Mach. Learn. Res. **10**, 1391–1445 (2009)
7. Kanamori, T., Suzuki, T., Sugiyama, M.: Statistical analysis of kernel-based least-squares density-ratio estimation. Machine Learning **86**(3), 335–367 (2012)
8. Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F.: A unifying view on dataset shift in classification. Pattern Recognition **45**(1), 521–530 (2012)

9. Nguyen, X., Wainwright, M.J., Jordan, M.I.: Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory **56**(11), 5847–5861 (2010)
10. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference **90**(2), 227–244 (2000)
11. Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., Kawanabe, M.: Direct importance estimation with model selection and its application to covariate shift adaptation. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) NIPS (2007)
12. Wen, J., Yu, C.-N., Greiner, R.: Robust learning under uncertain test distributions: relating covariate shift to model misspecification. In: Jebara, T., Xing, E.P. (eds.) Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pp. 631–639 (2014)
13. Zadrozny, B.: Learning and evaluating classifiers under sample selection bias. In: Greiner, R., Schuurmans, D. (eds.) Proceedings of the 21st International Conference on Machine Learning (ICML 2004) (2004)