ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

# Reducing Variance due to Importance Weighting in Covariate Shift Bias Correction

Van-Tinh Tran[1] and Alex Aussem[1]

1- University of Lyon 1, LIRIS, UMR 5205
69622 Lyon, France

**Abstract**. Covariate shift is a problem in machine learning when the input distributions of training and test data are different ($p(x) \neq p'(x)$) while their conditional distribution $p(y|x)$ is 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, which is harmful when the initial training sample size is already small. In this paper, we show that there exists a weighting scheme on the unlabeled data such that the combination of the weighted unlabeled data and the labeled training data mimics the test distribution. We further prove that the labels are missing at random in this combined data set and thus can be imputed safely in order to mitigate the undesirable sample-size-reduction effect of importance weighting. A series of experiments on synthetic and real-world data are conducted to demonstrate the efficiency of our approach.

## 1 Introduction

Covariate shift bias [5] occurs when samples are preferentially selected to the training data set, depending on the values of the input features, i.e. the marginal distribution of the input features in the training data, $p(x)$, differ from that of the test data, $p'(x)$, while the conditional distribution, $p(y|x) = p'(y|x)$, is the same. One may account for the difference between $p(x)$ and $p'(x)$ by reweighting the training examples using the so-called importance weight, denoted as $\beta(x) = p'(x)/p(x)$. A large body of research has been devoted to the estimation of this importance weight [1, 2]. However, reweighting methods do not necessarily improve the prediction accuracy as they reduce the effective training sample size. This becomes a severe problem when the initial training sample size is small. Another drawback of current importance weighting approaches is that the unlabeled data set are usually discarded once the importance weights are estimated. Some information is lost in the process. In [6] for instance, a combination of the weighted and the unweighted models is used to reduce the model variance, however the unlabeled cases are not used in the construction of the supervised model.

Covariate shift may be seen as a special case of sample selection bias, where data are collected through a binary variable $S$ that controls the selection of cases in the training set. We only have access to the cases for which $S = 1$. In this paper we show that there exists a weighting scheme on the unlabeled data so that a combination of these weighted unlabeled data and original training data forms

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

a new data set, called the *hybrid data set*, that have label missing at random (MAR). The missing values of label in the hybrid data are then imputed using state of the art imputation methods for MAR data. This approach is particularly useful when very few labeled data are provided.

## 2 The Hybrid Data Method

Assuming that the unlabeled data follow the input distribution $p'(x)$ of test data, we first derive, in this Section, a weighting scheme $w(x)$ on the unlabeled data so that a combination of these weighted unlabeled data and the original training data forms a new data set that mimics $p'(x)$. Our main result can be stated as follows:

**Theorem 1.** *Given $n_1$ training examples and $n_2$ unlabeled examples, that follow distributions $p(x)$ and $p'(x)$ accordingly, there exists a weighting scheme*

$$w(x) = \frac{n_1}{n_2}\left(\max_{x\in\mathcal{X}} \frac{p(x)}{p'(x)} - \frac{p(x)}{p'(x)}\right)$$

*on the unlabeled examples such that the mixture of $n_1$ unweighted training examples and $n_2$ weighted unlabeled examples follows the distribution $p'(x)$.*

*Proof.* The hybrid data set follows a mixture distribution which is:

$$p(x)\frac{n_1}{n_1 + n_2\int w(x)p'(x)dx} + p'(x)\frac{w(x)}{\int w(x)p'(x)dx} \times \frac{n_2\int w(x)p'(x)dx}{n_1 + n_2\int w(x)p'(x)dx}$$

By replacing $w(x)$ into the expression, we obtain $p'(x)$ □

We have shown that the resulting hybrid data set is unbiased but it still contains missing labels. There are circumstances under which even the best designed study is jeopardized by non-missing-at-random data. The following result shows the labels are in fact MAR:

**Theorem 2.** *The labels in the hybrid data set obtained from the weighting scheme in Theorem 1 are missing at random.*

*Proof.* From Theorem 1, the hybrid data set follows the marginal distribution $p'(x)$ of the test data. In addition, because of the definition of covariate shift, Let $R_Y = 1$ denotes "$Y$ is missing" and 0 otherwise, it is easily shown that $p(y|x, R_Y = 1) = p(y|x, R_Y = 0) = p(y|x)$, which is the definition of the MAR missing mechanism. □

Missing data imputation is a well-studied topic in statistical analysis. From the many references, we choose Predictive Mean Matching (PMM), which was first presented in [3] and proved to be successful with missing data imputation, as was shown to be robust to the misspecification of the imputation model in [4]. For the covariate shift problem, if we can choose a correctly specified model in

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

the first place, there will be no learning bias. However due to the lack of domain knowledge, it is safer to assume that the imputation model for the unlabeled data is misspecified. Robustness of imputation models to misspecification is an important criterion that should be considered with great care when choosing an imputation method. The methods for correcting covariate shift bears similarity to the techniques employed in semi-supervised learning. The latter usually make further assumptions on the data distribution $p$, more specifically on the relationship between $p(y|x)$ and $p(y)$ [7].

## 3 Experiments

In this section, we assess the ability of our hybrid data approach to reduce the model variance due to importance weighting in the covariate shift bias reduction process. We use two strategies to estimate the importance weights $\beta(x) = \frac{p'(x)}{p(x)}$: the first is based explicitly on the true bias mechanism, the second is based on Unconstrained Least-Square Importance Fitting (uLSIF). From the many references, we choose uLSIF estimator since it was proved to be successful with covariate shift. We first study a toy regression problem to show whether covariate shift corrections based on our method can reduce the prediction error on the test set when the learning model is misspecified and the training sample size is small. Then we test our approach on real world benchmark data sets corrupted by a simple covariate shift bias selection mechanism.

### 3.1 Toy regression problem

Consider the following training data generating process: $x \sim N(\mu_0, \sigma_0)$ and $y = f(x) + \epsilon$, where $\mu_0 = 0.5$, $\sigma_0 = 0.5$, $f(x) = -x + x^3$, and $\epsilon \sim N(0, 0.3)$. In the test data, the same relationship between $x$ and $y$ holds but the distribution of the covariate $x$ is shifted because of selection bias that causes examples to be selected with a probability depending on $x$:

$$p(s = 1|x) = \begin{cases} 4x^2 \text{ if } 4x^2 \in [0.01, 1] \\ 0.01 \text{ if } 4x^2 \leq 0.01 \\ 1 \text{ otherwise.} \end{cases}$$

The training and test distributions, along with their ratio are plotted in Fig. 1a and 1b. Least Square Regression is used to learn a linear model to predict output $y$ from $x$. We first investigate the effect of number of unlabeled data on the hybrid data approach performance. As may be seen in Figure 1c, the Mean Square Error (MSE) of the regression model drops as the unlabeled-labeled sample size ratio, $n_2/n_1$, increases. At first, as more unlabeled data are used, $n_2/n_1$ varies from 0 to 1, and the improvement is clearly noticeable. The smaller the initial training sample size, the larger the margin of the improvement as the hybrid data approach is more effective at preserving the effective sample size. When $n_2/n_1$ varies from 1 to 2 a further but moderate improvement is observed.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Again, the more unlabeled data are used, the smaller the weights of the unlabeled example according to Theorem 1. Consequently, the imputation variance contributes less to the final prediction error. Finally, when greater values of $n_2/n_1$, no improvement is noticed since the unlabeled data are only helpful in reducing the distribution mismatch up to the point when the hybrid data mimics closely the test data distribution. In contrast, the predictive performance of semi-supervised learning methods tend to increase as more unlabeled data are used. We will use in the toy problem an unlabeled data set five times larger than the labeled data set for and only twice as large in real-world data set experiments. We shall now compare the "hybrid-data approach" against respectively the unweighted, weighted, and hybrid-model approaches. In the hybrid-model approach presented in [6], the predictive performance in some regions of the input space is improved by combining the weighted and the unweighted models. The average MSE of these models over 100 repeated trials is reported for every training sample size in Figure 1d. The unweighted model (black solid line) serves as a baseline. As expected, it performs worse than the other models. When the training sample size is large enough (say, more than 300) the hybrid-model method achieves a lower MSE because it has the lowest bias as suggested by Theorem 2 in [6]. On the other hand, the hybrid-data method (blue solid line) outperforms any other method with a large margin when training sample size is small. As sample size increases, the variance reduction becomes less significant, the hybrid data's performance is similar to that of the weighted model. From these observations, we may conclude that the hybrid-data approach is more effective when the sample size is small.

### 3.2 Experiments on Real-world Datasets

In this experiment, we consider learning problems under a covariate shift induced by an artificial selection mechanism with known or estimated selection probabilities. We apply this selection scheme on a variety of UCI data sets in order to assess the efficiency of our approach in more realistic scenarios. We use a SVM classifier for both classification and regression tasks. Experiments are repeated 50 times for each data set. In each trial, we randomly select 100 training examples, 200 unlabeled examples, and an input feature $x^c$ that controls the probability of an an example to be selected in the training set as follows:

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

where $\sigma(x^c)$ denotes the standard deviation of $x^c$. Each of three approaches, namely the weighted data, hybrid model, and hybrid data is applied with both the true important weights and the important weights estimated with uLSIF. The MSE of each model is normalized by that of the unweighted model (our gold standard) and plotted in Fig.2a and 2b. As may be observed, the hybrid

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

data approach always outperforms the weighted model by a noticeable margin except when ulSIF is used on the Cadata data set. However, we suspect that the estimation of importance ratio on this data set fails as all other methods using ulSIF performs worse than the basic unweighted method on this data set. The hybrid data method also outperforms the hybrid model method in most situations, except on the Australian credit data set with true important weight and on the Cadata and Ionosphere data sets with ulSIF. Our results strongly suggest that our bias correction method combined with missing at random label imputation is effective at increasing the prediction performance when few labeled data are available.

## 4 Conclusions

In this paper, we presented a weighting scheme on the unlabeled data such that a combination of these weighted unlabeled data and the original training data forms a new data set that mimics the test distribution. The covariate shift was then formulated as a classification problem with labels missing at random in order to mitigate the sample-size-reduction effect of importance weighting. The efficiency of this bias reduction approach was demonstrated on both synthetic and real-world data.

## References

[1] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *NIPS*, pages 601–608. MIT Press, 2006.

[2] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.*, 10:1391–1445, Dec. 2009.

[3] R. J. Little. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296, 1988.

[4] T. P. Morris, I. R. White, and P. Royston. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology*, 14(1):1, 2014.

[5] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, Oct. 2000.

[6] V. Tran and A. Aussem. A practical approach to reduce the learning bias under covariate shift. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II*, pages 71–86, 2015.

[7] X. Zhu. Semi-supervised learning literature survey. Technical report, WisconsinMadison, 2005.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

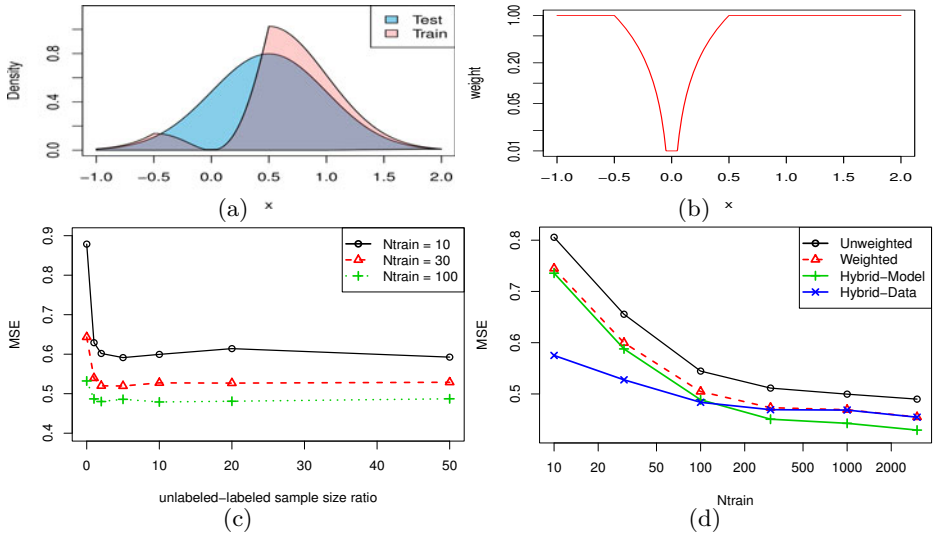
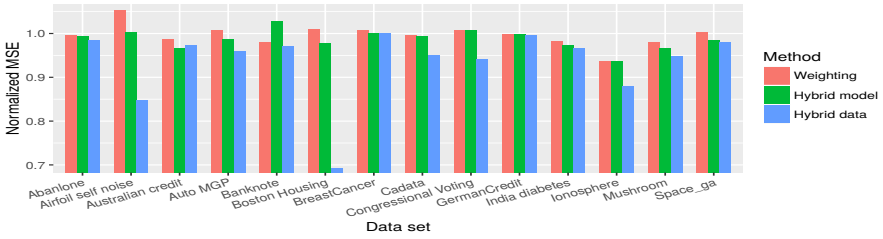
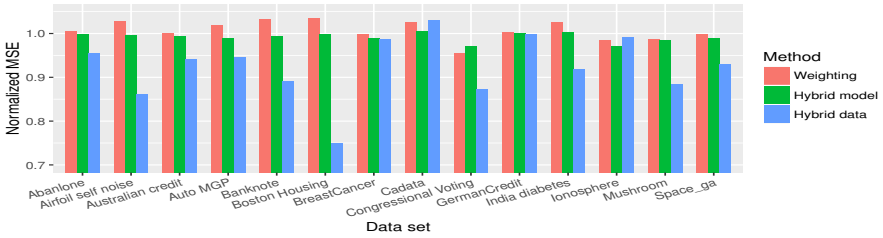



Fig. 1: A function f(x) is fitted by a linear model: a) Input density distribution; b) True importance weights; c) MSE of hybrid-data model vs. unlabeled/labeled ratio for different training sample sizes; d) MSE vs training sample size (on log scale) with unweighted data, weighted data, hybrid model, and hybrid data.




(a) The importance weights are derived from true selection probability.




(b) The importance weights are estimated using ulSIF.

Fig. 2: MSE gain of the weighted hybrid model (over the unweighted model) and the hybrid data method on each real-world data set.