Correcting a Class of Complete Selection Bias with External Data Based on Importance Weight Estimation

Van-Tinh Tran, Alex Aussem

Correcting a Class of Complete Selection Bias with External Data based on Importance Weight Estimation

Van-Tinh Tran and Alex Aussem

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

Abstract. We present a practical bias correction method for classifier and regression models learning under a general class of selection bias. The method hinges on two assumptions: 1) a feature vector, X_s, exists such that S, the variable that controls the inclusion of the samples in the training set, is conditionally independent of (X, Y) given X_s; 2) one has access to some external samples drawn from the population as a whole in order to approximate the unbiased distribution of X_s. This general framework includes covariate shift and prior probability shift as special cases. We first show how importance weighting can remove this bias. We also discuss the case where our key assumption about X_s is not valid and where X_S is only partially observed in the test set. Experimental results on synthetic and real-world data demonstrate that our method works well in practice.

Keywords: Selection bias, importance weighting, graphical models

1 Introduction

Selection bias, which occurs when training and test joint distributions are different, i.e. $P_{tr}(x, y) \neq P_{te}(x, y)$, is pervasive in almost all empirical studies, including Machine Learning, Statistics, Social Sciences, Economics, Bioinformatics, Biostatistics, Epidemiology, Medicine, etc. It is therefore highly desirable to devise algorithms that remain effective under such distribution shifts. In general, the estimation problem with two different distributions $P_{tr}(x, y)$ and $P_{te}(x, y)$ is unsolvable, as the two terms could be arbitrarily far apart. However, when $P_{tr}(x, y)$ and $P_{te}(x, y)$ differ only in $P_{tr}(x)$ and $P_{te}(x)$ (known as *covariate shift*) or only in $P_{tr}(y)$ and $P_{te}(y)$ (known as *prior probability shift*), effective adaptation is possible. In this paper, we present a practical correction method for classifier and regression model learning under a more general class of selection bias. We assume implicitly that there exists a joint probability distribution $P(x, y, s)$ that satisfies: $P_{te}(x, y) = P(x, y) = \sum_s P(x, y, s)$ and $P_{tr}(x, y) = P(x, y|s = 1)$, where

the variable S controls the selection of examples in the training set (1 means the example is selected, 0 means the example is not selected). While $P_{te}(x, y)$ and $P_{tr}(x, y)$ are derived from the same distribution $P(x, y, s)$, we assume no independence assumption holds between X, Y, and S. This is termed *complete selection bias* in the literature. In this case, we have to resort to some additional information on the mechanism by which the samples were preferentially selected to the data set to correct the bias.

The recent paper by Bareinboin et al. [1] has been very influential in our thinking. Mirroring their work, we show that, if we have a combination of biased data and unbiased data and qualitative probabilistic assumptions that are deemed plausible about our sampling mechanism, our problem becomes solvable. More specifically, we assume we have access to a S-control feature vector, X_s, and some additional sample of the form (x_s) that is drawn from the population as a whole, such that S is conditionally independent of (X, Y) given X_s. Despite being limited to specific or idealized situations, this framework includes covariate shift and prior probability shift as special cases. We also consider the case where X_s is not fully measured in the target population. This situation typically arises in various clinical studies or epidemiological scenarios, where some variables are too difficult or costly to measure in the target population.

We show that one may account for the difference between $P_{tr}(x, y)$ and $P_{te}(x, y)$ by reweighting the training points using the so-called importance weight, denoted as $\beta(x_s)$. If the selection process is explicitly known, then $\beta(x_s)$ is simply given by $\frac{P(s=1)}{P(s=1|x_s)}$, otherwise, we resort to an external (bias-free) data set of X_s in order to estimate $\beta(x_s)$ directly [3, 5]. Furthermore, as one usually has a partial understanding of the sampling mechanism, we investigate whether covariate shift and prior probability shift corrections may help reduce complete selection bias despite not being valid.

2 Bias correction

In this section, we investigate the interplay between two types of variables, V_B and V_P, where V_B are variables collected under selection bias, $P(V_B|S = 1)$, and V_P are variables collected in the population-level, $P(V_P)$ [1]. We assume that $Y \in V_B$. In [1], Bareinbom et al. provide a sufficient condition for $P(x, y)$ to be recoverable when no data is gathered over X and Y in the population level. We extend slightly their result by considering also the case where either Y or some input variables in X are not only collected under selection bias, but also in the population-level, (i.e. $\{X, Y\} \cap V_P \cap V_B \neq \emptyset$),

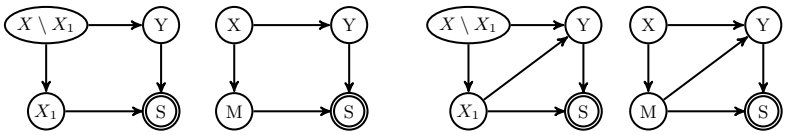
Theorem 1. *The bias-free distribution $P(x, y)$ is recoverable from a S-bias training samples if there exists a set of variables $X_s \subseteq V_B \cap V_P$, such that $S \perp\!\!\!\perp$*

[1] Upper-case letters in italics denote random variables (e.g., X, Y) and lower-case letters in italics denote their values (e.g., x, y). X denotes the input variables and Y the target.

$(X,Y)|X_s$ *and the support of* $P(x_s|s=1)$ *contains the support of* $P(x_s)$*. Let* $\beta(x_s) = \frac{P(s=1)}{P(s=1|x_s)}$*,* $P(x,y)$ *is then given by the formula,*

$$P(x,y) = \sum_{x_s \setminus \{x,y\}} P(x,y,x_s|s=1)\beta(x_s) \tag{1}$$

$\beta(x_s)$ can be reformulated as $\beta(x_s) = \frac{P(x_s)}{P(x_s|s=1)}$. So $\beta(x_s)$ may be estimated from a combination of biased and unbiased data. Theorem 1 relies on qualitative assumptions (X_s controls S over (X,Y)) that may appear difficult to satisfy in practice. However, in certain domains like epidemiology, information about the selection process can sometimes be expressed and modeled in a communicable scientific language (e.g., graphs or structural equations) by the domain experts. The selection bias mechanisms depicted in Figure 1 are common examples[2]. in epidemiology [2]. The directed acyclic graphs should be regarded as graphical structures encoding conditional independence between X, Y, and S which may involve other variables as as well variables, like M, that is not observed in the target domain, and thus that is not included as input variable to the model.



(a) $X_s = \{X_1, Y\}$ (b) $X_s = \{M, Y\}$ (c) $X_s = \{X_1, Y\}$ (d) $X_s = \{M, Y\}$

Fig. 1: Examples of complete selection bias mechanisms depicted graphically. The S-control vector is shown along each plot. X_1 is some input variable to the model, M denotes a variables that is not observed in the target domain, and thus not used a input variable.

Theorem 2. *Given that condition of Theorem 1 is satisfied, if* $\hat{P}$ *is a distribution such that:* $\hat{P}(x,y,x_s,s) = P(x,y,x_s,s)\beta(x_s)$ *then* $\hat{P}(x,y|s=1) \equiv P(x,y)$*.*

Proof.

$$\hat{P}(x,y,x_s,s=1) = P(x,y,x_s,s=1)\beta(x_s) = P(x,y,x_s,s=1)\frac{P(x_s)}{P(x_s|s=1)}$$

$$= P(s=1)P(x_s|s=1)P(x,y|x_s,s=1)\frac{P(x_s)}{P(x_s|s=1)} = P(s=1)P(x,y,x_s)$$

Thus, $\hat{P}(x,y,x_s,s=1) = P(x,y,x_s)P(s=1)$. If we sum this expression over x,y,x_s we obtain $\hat{P}(s=1) = P(s=1)$. Therefore,

$$\hat{P}(x,y,x_s|s=1) = \frac{\hat{P}(x,y,x_s,s=1)}{\hat{P}(s=1)} = \frac{P(x,y,x_s)P(s=1)}{P(s=1)} = P(x,y,x_s)$$

[2] We assume the reader is familiar with the concepts of d-separation [6]

Finally, $\hat{P}(x, y|s = 1) = \sum_{x_s \setminus \{x,y\}} \hat{P}(x, y, x_s|s = 1) = P(x, y)$. □

Theorem 2 states that an unbiased training sample can be obtained by weighting each training example by $\beta(x_s) = \frac{P(x_s)}{P(x_s|s=1)}$. Note however that the support of $P(x_s)$ should be contained in the support of $P(x_s|s = 1)$ for the $\beta(x_s)$ to be always defined. A similar technique applied to covariate shift was discussed in [9]. The unbiased expected loss of the model follows:

Corollary 1. *Given that condition of Theorem 1 is satisfied, and $\hat{P}$ in Theorem 2, for all classifier h, all loss function $l = l(h(x), y)$,*

$$E_{x,y\sim P}(l) = E_{x,y\sim \hat{P}}(l|s = 1)$$

$E_{x,y\sim P}(l)$ is the loss that we would like to minimize and $E_{x,y\sim \hat{P}}(l|s = 1))$ is the loss that may be estimated from the new biased sample drawn from weighted distribution $\hat{P}$.

3 Experiments

In this section, we assess the ability of importance weighting to remove complete selection bias based on Theorem 2. In the toy experiment, we investigate whether covariate shift and prior probability shift corrections may help reduce complete selection bias despite our assumptions between the training and test distributions difference being violated (through an invalid choice for X_s).

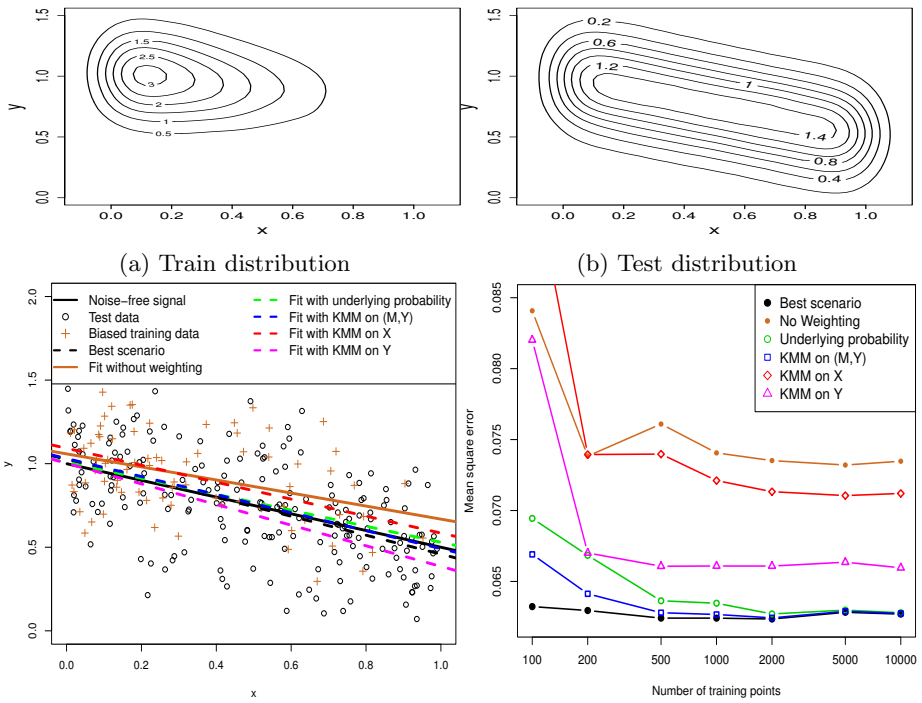
When the selection process is explicitly known, $\beta(x_s)$ is simply given by $\frac{P(s=1)}{P(s=1|x_s)}$. Otherwise, we resort to an external (bias-free) data set of X_s in order to estimate $\beta(x_s)$ as $\frac{P(x_s)}{P(x_s|s=1)}$. In this study, we use the Kernel Mean Matching (KMM) [3, 8] estimator for $\beta(x_s)$ denoted as KMM(X_s). As one usually has a partial understanding of the sampling mechanism, we investigate whether covariate shift (i.e., $\beta(x) = \frac{P(x)}{P(x|s=1)}$) and prior probability shift (i.e., $\beta(y) = \frac{P(y)}{P(y|s=1)}$), corrections may help reduce complete selection bias despite not being valid. These strategies are denoted as KMM(X), KMM(Y). We first apply our method to a simple toy problem and then compare KMM to another estimator called the Unconstrained Least-Square Importance Fitting (uLSIF) [5] on a variety of regression and classification benchmarks from the UCI Archive.

3.1 Toy problem

Consider the S-bias mechanism displayed in Fig. 1b, where the feature X has a uniform distribution in $[0, 1]$: $P(X) \sim \mathcal{U}(0, 1)$. Note that the influence of M on Y is mediated by $\{X, S\}$. The observations are generated according to $y = 1-0.5x$ and are observed in Gaussian noise with standard deviation 0.5 (see Fig. 2c; the black solid line is the noise-free signal). The intermediate variable M, between X and S, is generated according to $M = X+\mathcal{N}(0, 0.3^2)$. As M is only measured in the training set, it is not used as an input variable in our regression model.

Therefore, we investigate a case where X_s is partially missing in the test set. The probability of a given example being included in the training set depends on Y and M and is given by

$$P(S=1|m,y) \sim \begin{cases} y-m, & \text{if } 0.1 \leq (y-m) \leq 1 \\ 0.1, & \text{if } (y-m) \leq 0.1 \\ 1, & \text{otherwise} \end{cases}$$



(a) Train distribution (b) Test distribution

(c) Linear model fit (d) Mean square error

Fig. 2: Toy regression problem 1. (a) and (b) Contour plots X-Y on training and test sets; (c) Polynomial models of degree 1 fit with OLS and WOLS; (d) Average performances of four WOLS methods and OLS on the test data as a function of the number of training points.

Note that the minimum value of $P(S=1|m,y)$ needs to be greater than 0 so that the support of $P(m,y)$ is contained in the support of $P(m,y|s=1)$, as required by Theorem 1. The choice of $P(m,y)$ is intended to induce a noticeable discrepancy between $P(y|x,s=1)$ and $P(y|x)$. We sampled 200 training (red crosses in Fig. 2c) and testing (grey circles) points from P_{tr} and P_{te} respectively. The bias is clearly noticeable from the X-Y contour plots in Fig. 2a and b. The

bias-free distribution $P(x, y)$ is recoverable from the S-bias training samples since $\{M, Y\}$ satisfies Theorem 1. Thus we use Theorem 2, to remove selection bias by weighting each example by the importance ratio:

$$\beta(x_s) = \beta(m, y) = \frac{P(m, y)}{P(m, y|s = 1)} = \frac{P(s = 1|m, y)}{P(s = 1)}$$

where $P(s = 1|m, y)$ and $P(s = 1)$ may be obtained from the known selection mechanism shown above or directly estimated by KMM. We attempted to model the observations with a degree 1 polynomial. The black dashed line in Fig. 2c is a best-case scenario given our test points, which is shown for reference purposes: it represents the model fit using ordinary least squared (OLS) on the test set. The brown line is a second reference result, derived only from the training data via OLS, and predicts the test data very poorly. The green dashed line is a third reference result, fit with weighted ordinary least square (WOLS), using the true $\beta(x_s)$ values calculated from the true data generating mechanism, and predicts the test data quite well. The other three dashed lines are fit with WOLS using the KMM weighting schemes under the three assumptions. Note that the true generating model between X and Y is included in the hypothesis space. We estimated the effect of the number of training points on the estimation of the reweighting factors by examining the average mean square error (MSE) on the test set as a function of the number of training points. As may be observed in Fig. 2d, the error goes down as the sample size increases, until it reaches an asymptotic value. KMM(X_s) performs well even with relatively moderate amounts of data achieving almost optimal error quite quickly, handily outperforming the reweighting method based on KMM(X) and KMM(Y) by a noticeable margin. More interestingly, KMM(X_s) also outperforms the reweighting method based on the true data generating mechanism, especially when sample size is small. This result may seem counter-intuitive at first sight: the reason is that the exact importance-sampler weights are not always optimal unless we have an infinite sample size. See [7] for a thorough discussion. Remarkably, despite our assumption regarding the difference between the training and test distributions being violated, KMM(Y) and KMM(X) improve the test performance.

3.2 Real-world data sets

We now examine whether using importance weighting can reduce selection bias in 10 UCI data sets with 5 classification tasks and 5 regression tasks. We employ three methods to estimate importance weighting: ratio of underlying probability, KMM and uLSIF and compare their performance against the baseline unweighted method. For each data set, X_s is chosen to be the label Y and the most correlated input variable to Y (denoted as X_1 for simplicity). The selection bias mechanism is illustrated in Fig. 1c. The selection variable S for each training example is determined according to two scenarios depending on whether it is regression or classification problem. For regression problem, we use $P(s = 1|x_1, y) = exp(ax_1 + by + c)/[1 + exp(ax_1 + by + c)]$, where a, b, c, are parameters that determine the bias. For binary classification problem, we use:

$$P(s=1|x_1,y)=\begin{cases}0.5 & \text{if } x_1 > mean(x_1) \text{ and } y=1\\ 1 & \text{if } otherwise.\end{cases}$$

For each data set, we then train 4 predictive models learned under the four weighting schemes discussed above and a model learned from the unbiased data (baseline) using SVM-light [4] which allows importance weighting to be fed directly to SVM. All classifiers are trained with the common Radial Basis Function (RBF), with a kernel size σ chosen through a 5-fold cross validation. This procedure is repeated 100 times for each data set.

Data set	No weighting	KMM	uLSIF	Underlying P	Unbiased model
India diabetes	0.338 ± 0.049	0.266 ± 0.040	0.332 ± 0.053	0.287 ± 0.055	0.258 ± 0.035
Ionosphere	0.069 ± 0.039	0.066 ± 0.039	0.067 ± 0.040	0.067 ± 0.039	0.065 ± 0.036
BreastCancer	0.044 ± 0.016	0.039 ± 0.015	0.043 ± 0.017	0.040 ± 0.016	0.038 ± 0.015
Haberman	0.264 ± 0.069	0.262 ± 0.071	0.263 ± 0.070	0.262 ± 0.071	0.262 ± 0.071
GermanCredit	0.300 ± 0.044	0.298 ± 0.046	0.298 ± 0.045	0.298 ± 0.046	0.295 ± 0.046
Airfoil self noise*	0.534 ± 0.104	0.470 ± 0.122	0.475 ± 0.082	0.445 ± 0.081	0.403 ± 0.059
Abanlone*	0.526 ± 0.048	0.484 ± 0.054	0.521 ± 0.057	0.466 ± 0.041	0.456 ± 0.036
Computer Hardware*	0.326 ± 0.308	0.321 ± 0.304	0.321 ± 0.299	0.319 ± 0.307	0.305 ± 0.201
Auto MGP*	0.268 ± 0.148	0.298 ± 0.192	0.212 ± 0.129	0.203 ± 0.128	0.129 ± 0.063
Boston Housing*	0.323 ± 0.110	0.327 ± 0.127	0.349 ± 0.133	0.332 ± 0.127	0.298 ± 0.112

Table 1: Mean test error averaged over 100 trials of different weighting schemes on UCI data set. Data sets marked with * are for regression problems

The results are reported in Tab. 1. As may be seen, all importance weighting schemes achieve lower prediction error with respect to the baseline unweighted scheme. The underlying probability weighting scheme performs pretty good. Curiously, on the Boston Housing data set, all three weighting schemes perform worse than the baseline unweighted method. It seems therefore that the effectiveness of bias correction based on importance weighting is data dependent. In order to better assess the overall results obtained for each of the 4 weighting schemes, a non-parametric Friedman test was firstly used to evaluate the rejection of the hypothesis that all the models perform equally well (except the unbiased model of course) at significant level 5%. Statistically significant differences were observed. So we proceeded with the Nemenyi post hoc test. The results along with the average rank diagrams are shown in Fig. 3. The ranks are depicted on the axis, in such a manner that the best ranking algorithms are at the rightmost side of the diagram. The algorithms that do not differ significantly (at $p = 0.05$) are connected with a line. As may be observed in Fig. 3, contrary to uLSIF, KMM is significantly better than no weighting.

4 Discussion & Conclusion

The aim of this paper was to elaborate on the idea of exploiting the assumptions that are deemed plausible about the sampling mechanism to correct or reduce

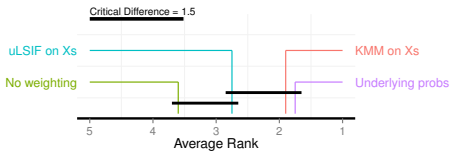


Fig. 3: post hoc analysis

selection bias in machine learning tasks. The method hinges on the existence of a S-control feature vector, X_s, and an additional (biased-free) sample that allows us to estimate the distribution of X_s. We showed experimentally that direct weighting estimation is able to achieve significant improvements in accuracy over the unweighted method, even in situations where our key assumption is not valid (assuming covariate shift and prior probability shift instead of complete selection bias). However the gain in accuracy is data dependent. In fact, all conclusions are extremely sensitive to which variables we choose for X_s. As the choice of X_s usually reflects the investigator's subjective and qualitative knowledge of statistical influences in the domain, the data analyst must weight the benefit of reducing selection bias against the risk of introducing new bias carried by unmeasured covariates even where none existed before. Nevertheless, we hope this study will convince others about the importance of selection bias correction methods in practical studies and suggest relevant tools which can be used to achieve that goal.

References

1. Bareinboim, E., Tian, J., Pearl, J.: Recovering from selection bias in causal and statistical inference. In: Brodley, C.E., Stone, P. (eds.) AAAI (2014)
2. Hernán, M.A., Hernández-Díaz, S., Robins, J.M.: A structural approach to selection bias. Epidemiology 15(5), 615–625 (2004)
3. Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., Schölkopf, B.: Correcting sample selection bias by unlabeled data. In: Schölkopf, B., Platt, J.C., Hoffman, T. (eds.) NIPS. pp. 601–608. MIT Press (2006)
4. Joachims, T.: Making large-scale svm learning practical. LS8-Report 24, Universität Dortmund, LS VIII-Report (1998)
5. Kanamori, T., Hido, S., Sugiyama, M.: A least-squares approach to direct importance estimation. J. Mach. Learn. Res. 10, 1391–1445 (2009)
6. Pearl, J.: Probabilistic reasoning in intelligent systems - networks of plausible inference. Morgan Kaufmann series in representation and reasoning, Morgan Kaufmann (1989)
7. 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)
8. Yu, Y., Szepesvári, C.: Analysis of kernel mean matching under covariate shift. In: Langford, J., Pineau, J. (eds.) ICML (2012)
9. Zadrozny, B.: Learning and evaluating classifiers under sample selection bias. In: Brodley, C.E. (ed.) ICML. pp. 114– (2004)