Fair Transfer Learning with Missing Protected Attributes

Amanda Coston,[1,2] Karthikeyan Natesan Ramamurthy,[1] Dennis Wei,[1] Kush R. Varshney,[1]
Skyler Speakman,[3] Zairah Mustahsan,[4] and Supriyo Chakraborty[1]

[1]IBM Research, Yorktown Heights, NY, USA, [2]Carnegie Mellon University, Pittsburgh, PA, USA
[3]IBM Research, Nairobi, Kenya, [4]IBM Watson AI Platform, Yorktown Heights, NY, USA

ABSTRACT

Risk assessment is a growing use for machine learning models. When used in high-stakes applications, especially ones regulated by anti-discrimination laws or governed by societal norms for fairness, it is important to ensure that learned models do not propagate and scale any biases that may exist in training data. In this paper, we add on an additional challenge beyond fairness: unsupervised domain adaptation to covariate shift between a source and target distribution. Motivated by the real-world problem of risk assessment in new markets for health insurance in the United States and mobile money-based loans in East Africa, we provide a precise formulation of the machine learning with covariate shift and score parity problem. Our formulation focuses on situations in which protected attributes are not available in either the source or target domain. We propose two new weighting methods: prevalence-constrained covariate shift (PCCS) which does not require protected attributes in the target domain and target-fair covariate shift (TFCS) which does not require protected attributes in the source domain. We empirically demonstrate their efficacy in two applications.

KEYWORDS

Fairness, transfer learning, risk assessments

ACM Reference Format:
Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair Transfer Learning with Missing Protected Attributes. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3306618.3314236

1 INTRODUCTION

The covariate shift setting in machine learning is often encountered in real-world applications that have limitations on data collection and require training on a different probability distribution than the one a model will ultimately be tested on [3, 20, 23]. In this setting, the training (source) and test (target) marginal feature distributions are different but the conditional distribution of labels given features is the same in the training and test distributions. Unsupervised domain adaptation to account for covariate shift can be achieved by forms of transfer learning [11, 26]. Relevant applications include risk assessment of people for credit and insurance when the provider has historical features and risk labels about its members, but is expanding into new markets with different demographics for which it does not have risk labels [22, 28]. The provider must learn models from its existing market data (source distribution) to score people in the new market (target distribution).

In this paper, we propose the problem of fair transfer learning under covariate shift and methods of solution. We examine the variations of the problem in which the protected attributes are available only for the source or only available for the target. We develop two new weighting methods: prevalence-constrained covariate shift (PCCS) which does not require protected attributes in the target domain and target-fair covariate shift (TFCS) which does not require protected attributes in the source domain. Weighting methods are commonly used for fairness and covariate shift, and they have the advantage of being compatible with many classification methods.

The issue of missing protected attributes is not only encountered in covariate shift settings, but is a more general challenge [18, J. Langford in panel discussion] because legal restrictions often prevent the collection of protected attributes or their joining to the rest of the dataset. For example, under Title VII of the 1964 Civil Rights Act, employers cannot ask potential applicants about gender [4]. Gupta et al. [13] address the problem of unavailable protected attributes by constructing proxy groups using variables in the dataset that are not protected but are likely correlated to protected attributes based on prior subject matter knowledge. Our approach is neither to explicitly construct proxy groups nor to use other variables, but to use the protected attributes themselves in related datasets. This approach enables companies to audit their hiring practices for gender discrimination by using aggregated datasets like the American Community Survey that have gender as a feature. Financial institutions could also use this to evaluate whether their credit approval processes comply with regulations prohibiting discrimination in cases where they are legally or practically unable to collect the protected attribute [5].

We apply our method to the medical expenditure dataset produced by the US Department of Health and Human Services known as the Medical Expenditure Panel Survey (MEPS). Health insurance providers in the US can choose the markets in which they will offer their plans and in which ones they will not. This decision making is driven by a risk assessment of the market, and thus the task is predicting individuals with low total healthcare expenditure vs. high expenditure over the course of a year. Section 1557 of the Patient Protection and Affordable Care Act (PPACA) in the United States made discrimination in healthcare on the basis of sex and race illegal [10, 14, 27]. As the PPACA was enacted, many insurance companies expanded rapidly into new markets, encountering the need to perform risk assessment having expenditure data only from

AIES '19, January 27–28, 2019, Honolulu, HI, USA

ACM ISBN 978-1-4503-6324-2/19/01…$15.00
https://doi.org/10.1145/3306618.3314236

the markets they served, not from new markets. Wei et al. [28] presented a covariate shift-based solution to this problem without considering fairness.

We also evaluate our methods on mobile money loan approvals in East Africa. Based on the success of mobile phone-based savings products such as M-Pesa, providers have recently begun offering credit services [7]. Algorithms make loan approval decisions based on mobile usage data. There has been rapid expansion by financial service providers into new markets (in this case, different countries with different partnering mobile network operators) with a need for covariate shift-based machine learning; one such solution is presented in [22]. We evaluate PCCS and TFCS on a dataset modeled after the actual data of a commercial financial institution in Africa.

2 PROBLEM SETTING

2.1 Covariate Shift

We are given labeled data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ from a source domain where $y_i \in \{0, 1\}$. We assume one of the labels is a more favorable outcome than the other. We also have unlabeled data $\{x_{n+1}, \ldots, x_{n+m}\}$ from a target domain to which we wish to assign labels. We use $\mathcal{S} = \{1, \ldots, n\}$ and $\mathcal{T} = \{n+1, \ldots, n+m\}$ to distinguish the index sets from the source and target domains. With covariate shift, the features $x_i \in \mathbb{R}^d$, $i \in \mathcal{S}$, are assumed to be drawn from a source distribution with density $p_X(x)$ while x_i for $i \in \mathcal{T}$ are drawn from a different target distribution with density $q_X(x)$. It is assumed that the conditional distribution of Y, i.e. $p_{Y \mid X}(y \mid x)$, is the same in both domains.

The standard approach to supervised learning is to find a predictor $\hat{y}(x)$ in a class $\mathcal{H}$ that minimizes empirical risk,

$$\min_{\hat{y} \in \mathcal{H}} \frac{1}{n} \sum_{i \in \mathcal{S}} \mathcal{L}(\hat{y}(x_i), y_i), \tag{1}$$

where $\mathcal{L}$ is the loss function between $\hat{y}$ and y that defines the risk. As $n \to \infty$, the empirical risk converges to the population risk, which can be written as the iterated expectation

$$\mathbb{E}_{p_X}[\mathbb{E}[\mathcal{L}(\hat{y}(X), Y) \mid X]] \tag{2}$$

to emphasize that the outer expectation is with respect to the source distribution p_X. In this ideal limit where $\mathcal{H}$ also contains arbitrarily complex functions, the optimal predictor in both domains, $p_{Y \mid X}(\cdot \mid x)$, can be recovered wherever $p_X(x)$ is positive. However for finite samples and constrained $\mathcal{H}$, the predictor obtained by minimizing empirical risk (1) generally shows traces of p_X and may not be best suited to the distribution q_X under which testing occurs.

Many methods that address the covariate shift problem do so by weighting training/source instances with weights $w_i \geq 0$ so that (1) becomes

$$\min_{\hat{y} \in \mathcal{H}} \frac{1}{n} \sum_{i \in \mathcal{S}} w_i \mathcal{L}(\hat{y}(x_i), y_i).$$

If $w_i = q_X(x_i)/p_X(x_i)$, then the weighted empirical risk converges to (2) with p_X replaced by q_X, as is desired for the target domain. Multivariate density estimation is difficult and therefore it is common to estimate the ratio of densities directly. One approach treats this as a classification problem [3] where the label is the source or target distribution. Logistic regression does this naturally and has been shown to be optimal (minimum asymptotic variance) for correctly specified models [19] but performs poorly for mis-specified models [25]. We use logistic regression in Section 4 for illustration purposes, noting that more sophisticated methods exist, e.g. [12, 24].

2.2 Protected Attribute Availability

We consider the problem of fairness with respect to protected groups. The definition of protected groups is assumed given and depends on the application context. Let $g_i \in \{0, 1, \ldots, G-1\}$ represent the group identity of instance i, which may be determined by one or more *protected attributes* such as race and gender. We pay particular attention to situations in which protected attribute data are available only in the source or target domain. In the former case, the source data consists of triplets $\{(g_i, x_i, y_i), i \in \mathcal{S}\}$ while the target data is $\{x_i, i \in \mathcal{T}\}$ as before. In the latter, the source data is $\{(x_i, y_i), i \in \mathcal{S}\}$ while the target data becomes $\{(g_i, x_i), i \in \mathcal{T}\}$. Based on the values of $\{g_i\}$, we define the partition of the source data into sets $\mathcal{S}_k = \{i \in \mathcal{S} : g_i = k\}$, i.e. source examples belonging to group k, for all $k = 0, \ldots, G-1$. The partition $\{\mathcal{T}_k, k = 0, \ldots, G-1\}$ of the target is defined similarly.

2.3 Fairness Metrics

We focus on notions of *demographic* or *statistical parity* that deal with the dependence of the classifier output on the protected attributes. We define a *score* $s(x)$ as a function that assigns to a feature vector x a number in $[0, 1]$ corresponding to the likelihood of a positive outcome $Y = 1$ given x. $s(x; \theta)$ denotes a score function parametrized by a vector of parameters θ. Viewing X as a random variable, a distribution is induced for the score $s(X)$ as well. We will say that a score satisfies *score parity* in the *strong* sense if $s(X)$ is statistically independent of the protected group variable G. This notion may be relaxed by requiring some distributional distance $D\big(p_{s(X) \mid G}(\cdot \mid k), p_{s(X) \mid G}(\cdot \mid l)\big)$ between scores conditioned on groups $k \neq l$ to be bounded by some constant $\delta > 0$.

Herein we focus on two weaker and more common definitions of score parity. The first is *mean* or *average* score parity,

$$\mathbb{E}[s(X) \mid G = k] = \mathbb{E}[s(X) \mid G = l] \quad \forall k, l \in \{0, \ldots, G-1\} \tag{3}$$

which is the definition of "statistical parity" in e.g. [8]. Mean score parity may also be relaxed by allowing small deviations, and it is this relaxed definition that is targeted by the methods proposed in Section 3. The second notion is *thresholded* score parity at threshold t: $\forall k, l \in \{0, \ldots, G-1\}$,

$$\Pr(s(X) > t \mid G = k) = \Pr(s(X) > t \mid G = l) \tag{4}$$

which is also called "statistical parity" in e.g. [6]. Thresholded score parity applies when the score is thresholded to yield a binary prediction. It is used in Section 4 as a second fairness metric to evaluate different methods. If a score satisfies thresholded parity for all thresholds $t \in [0, 1]$, then it also satisfies parity in the strong sense above. It is clear that strong score parity implies both mean and thresholded parity. Moreover, if approximate strong parity holds in that $D\big(p_{s(X) \mid G}(\cdot \mid k), p_{s(X) \mid G}(\cdot \mid l)\big)$ is small, then one expects the mean score disparity and thresholded score disparity to be small as well although the details depend on the distance measure D.

A quantity not involving but related to scores is *prevalence*, which describes the proportions of class labels. For binary Y, it is sufficient

to assess prevalence $\Pr(Y = 1)$ of the positive outcome in the entire dataset and prevalence $\Pr(Y = 1|G = k)$ for particular groups. Prevalence differences between groups are therefore a measure of bias in the dataset. Since scores are often designed to estimate either $p_{Y|X}$ or Y itself after thresholding, controlling group-specific prevalences is a way of encouraging mean score parity in the former case or thresholded score parity in the latter, provided that the score is an approximately unbiased estimator, $\mathbb{E}[s(X)] \approx \mathbb{E}[Y]$ or $\Pr(s(X) > t) \approx \Pr(Y = 1)$. The method discussed in Section 3.1 relies on this relationship between prevalences and scores.

3 PROPOSED METHODS

Given the popularity of weighting methods for the covariate shift problem and in works on fairness [1, 15, 17], we focus in this paper on weighting as a means to address covariate shift and fairness jointly. Our goal is to determine weights $w_i \geq 0$ for the source/training instances (x_i, y_i), $i \in \mathcal{S}$. We propose two methods for the cases in which protected attribute information is available only for the source or target populations respectively. The method of Section 3.1 assumes nothing more than the use of a classification algorithm that accepts weights as input. The method of Section 3.2 requires differentiability of the classification loss function. It is possible to relax this assumption, for example by using smooth approximations to the loss and second-order optimization techniques as in [16]. The derivatives can be evaluated in closed form, as we do for logistic regression, or using automatic differentiation [2].

3.1 Prevalence-Constrained Covariate Shift (PCCS)

For the scenario in which the protected attribute is available for the source population but not the target population, we propose a method that combines conventional covariate shift with weighting to bring group-specific prevalences closer together. As discussed in Section 2.3, differences in prevalences characterize dataset bias, and controlling this bias encourages score parity. Let $w_{CS}(x)$ be a *covariate shift weight*, i.e. an approximation to the ratio $q_X(x_i)/p_X(x_i)$, obtained through logistic regression or other methods [3, 12, 24, 25]. The goal is to learn weights w_i for each training example that are as close as possible to the covariate shift weights subject to constraints on weighted prevalences. The objective function is thus

$$\min_{w} \sum_{i \in \mathcal{S}} |w_i - w_{CS}(x_i)|. \tag{5}$$

Norms other than the ℓ_1 norm can also be used. The prevalence constraints for fairness enforce closeness between all pairs of groups:

$$\frac{\sum_{i \in \mathcal{S}_k : y_i = 1} w_i}{\sum_{i \in \mathcal{S}_k} w_i} \geq \frac{\sum_{i \in \mathcal{S}_l : y_i = 1} w_i}{\sum_{i \in \mathcal{S}_l} w_i} - \delta \qquad \forall k, l \in \{0, \ldots, G-1\}, \tag{6}$$

where the parameter δ trades off between differences in prevalences and deviation of w from w_{CS}. To make these constraints convex, we add equality constraints on the proportion of weight allocated to each group:

$$\sum_{i \in \mathcal{S}_k} w_i = c_k \sum_{i \in \mathcal{S}} w_i, \quad k \in \{0, \ldots, G-1\} \tag{7}$$

where

$$c_k = \frac{\sum_{i \in \mathcal{S}_k} w_{CS}(x_i)}{\sum_{i \in \mathcal{S}} w_{CS}(x_i)}, \tag{8}$$

i.e. we require the allocations to groups specified by the covariate shift weights to remain unchanged. Lastly we require weights to be non-negative:

$$w_i \geq 0, \quad i \in \mathcal{S}. \tag{9}$$

The optimization problem is to minimize the objective in (5) subject to constraints (6)–(9).

3.2 Target-Fair Covariate Shift (TFCS)

We now consider the scenario in which the protected attribute is available for the target but not the source. We may directly evaluate the score disparity of the classifier on the target and adjust the classifier to reduce the disparity. We assume that the classifier parameters are chosen to minimize the weighted empirical risk,

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i \in \mathcal{S}} w_i \mathcal{L}(s(x_i; \theta), y_i), \tag{10}$$

where $\mathcal{L}(s(x; \theta), y)$ is a twice differentiable function of θ as discussed and any regularizer is absorbed into $\mathcal{L}$. Thus the classifier can be adjusted by changing the weights w_i.

To measure score disparity, we introduce the following fairness loss that sums the squares of average score disparities over all pairs of groups $(\mathcal{T}_k, \mathcal{T}_l)$:

$$\mathcal{L}_f\left(s(\cdot; \hat{\theta})\right) = \sum_{k<l} \left(\frac{1}{|\mathcal{T}_k|} \sum_{i \in \mathcal{T}_k} s(x_i; \hat{\theta}) - \frac{1}{|\mathcal{T}_l|} \sum_{i \in \mathcal{T}_l} s(x_i; \hat{\theta}) \right)^2. \tag{11}$$

The weights w_i are chosen to minimize a linear combination of this fairness loss with a classification loss:

$$\min_{w} \frac{1}{n} \sum_{i \in \mathcal{S}} w_{CS}(x_i) \mathcal{L}(s(x_i; \hat{\theta}), y_i) + \lambda \mathcal{L}_f\left(s(\cdot; \hat{\theta})\right), \tag{12}$$

where $w_{CS}(x_i)$ are covariate shift weights as in Section 3.1 and are *fixed* (not to be confused with w). The first term in (12) thus approximates classification loss on the target population by weighting the source population, where labels are available. The fairness loss $\mathcal{L}_f$ is evaluated on the target, where the protected attribute is available. Both terms are explicit functions of the scores parametrized by $\hat{\theta}$; the notation $s(\cdot; \hat{\theta})$ emphasizes this dependence. The parameters $\hat{\theta}$ are a function of the optimization variables w_i through (10).

We propose to optimize (12) through gradient descent. The algorithm (see Algorithm 1) alternates between gradient updates to w to decrease the objective in (12) and solving (10) to obtain a new classifier from the updated w, which is then re-evaluated using (12). In our experiments, we use a constant step size η and terminate after a fixed number of iterations.

The combined loss (12) is generally not a convex function of the weights w_i. Hence different initializations may lead to different solutions. One choice is to initialize with covariate shift weights, $w = w_{CS}$. In the case of $\lambda = 0$ in (12) (i.e. only classification loss), $w = w_{CS}$ is a stationary point as will be shown at the end of Appendix A. Accordingly for small λ, $w = w_{CS}$ is expected to be

Algorithm 1 Target-Fair Covariate Shift (TFCS)

Input:
Data: labeled source $\{(x_i, y_i), i \in \mathcal{S}\}$, target with protected attribute $\{(x_i, g_i), i \in \mathcal{T}\}$
Parameters: trade-off λ, step size η
Estimate covariate shift weights w_{CS} from $\{x_i, i \in \mathcal{S} \cup \mathcal{T}\}$
$w \leftarrow w_{CS}$ **or** $w_i \leftarrow 1\ \forall i$ (uniform)
repeat
  Learn classifier parameters $\hat{\theta}$ given w (10)
  Evaluate combined loss (12)
  # Gradient computations:
  Compute $\nabla_{\hat{\theta}} \mathcal{L}_c$ (e.g. (19))
  Compute $\nabla_{\hat{\theta}} \mathcal{L}_f$ (e.g. (20))
  Compute $\partial\hat{\theta}/\partial w_i\ \forall i$ (15)(16)
  Compute $\nabla_w \mathcal{L}_t$ (14)
  # Gradient update
  $w \leftarrow w - \eta \nabla_w \mathcal{L}_t$
until stopping criterion is met
Output: Classifier parameters $\hat{\theta}$, weights w

near-stationary. A simpler alternative is to initialize with uniform weights $w_i = 1$.

For the scenario in which the protected attribute is available in both the source and target domains, we propose to combine the methods in Section 3.1 and this section. First, PCCS is used to achieve covariate shift and approximate score parity based on the protected attribute in the source. Then the PCCS weights are used to initialize the minimization in (12) to refine score parity in the target domain.

Since the combined loss (12) is an indirect function of w via (10), the calculation of its gradient with respect to w is non-standard. Appendix A derives the necessary expressions.

4 EXPERIMENTS

We demonstrate the utility of PCCS and TFCS in fair transfer learning for two applications. The first is a healthcare cost prediction scenario when a health insurance company that is servicing an existing market wants to venture into a new market, while ensuring equal benefit to all race- and gender-based intersectional groups in the new market. The second is a loan approval setting where a mobile money provider from one country in East Africa is expanding to another country while being non-discriminatory according to age and gender.

We use AUC to assess accuracy and we compute four fairness metrics:

(1) *Mean score parity (MSP) loss*: Square root of sum of squares of differences between mean scores for all pairs of groups (see equations (3) and (11)).
(2) *Thresholded score parity (TSP) loss*: Square root of sum of squares of differences between thresholded scores for all pairs of groups (see equation (4)).
(3) *Max* Δ *MSP*: Maximum of absolute differences between mean scores for all pairs of groups.
(4) *Max* Δ *TSP*: Maximum of absolute differences between thresholded scores for all pairs of groups.

We compare our proposed methods against four baselines:

(1) *Native*: Train and test on the target population (i.e. not in the transfer learning setting).
(2) *Unadapted transfer learning*: Train on the source population and test on target population without any adaptation to the target population during training.
(3) *Covariate shift*: Train on the source population reweighed to resemble the target population.
(4) *Kamiran Calders*: Correct the source dataset for fairness using the approach proposed in [15] without performing any covariate shift.

Testing is always performed on the target population where true labels and protected attributes are used to evaluate accuracy and fairness. PCCS and Kamiran-Calders use protected attributes only from the source while TFCS uses them only from the target.

In all cases, logistic regression is used as the classification algorithm and also to obtain covariate shift weights. This choice is intended as a simple illustration of the methods and richer models can certainly be substituted. The TFCS method was initialized using uniform weights and its stopping criterion is a maximum number of iterations, usually set to a few hundred, depending on the dataset used.

We studied the behavior of PCCS and TFCS for various values of their respective free parameters δ and λ. For future applications, the choice of the free parameter will depend on the particular setting (including the initial discrepancies in group prevalences) but generally we would recommend using a δ in $[0, .075]$ for PCCS. For TFCS, we recommend using cross-validation to choose λ.

We note that transfer learning can exacerbate discrimination or improve fairness; the direction depends on the application, and indeed, in our experiments we observe both. Regardless of whether transfer learning alone helps or hurts fairness, our fairness-aware transfer learning methods are able to improve the fairness metrics.

4.1 Medical Expenditure Panel Survey (MEPS)

The MEPS dataset [9] is obtained using a nationally representative survey of the US population. It contains annual healthcare cost, demographics, and self-reported medical information. We use the data from panel 19 of the 2015 survey. There is no concept of market in this dataset since it does not come from an insurance provider (those datasets are proprietary, but MEPS shares relevant characteristics with such datasets). We define the source market to consist of people earning less than national median income (USD 21,000), and target market to consist of the rest of the population.

We consider two protected attributes: gender and race. Both are legally protected, as discussed in Section 1. In our experiments, the races considered are non-Hispanic whites and non-Hispanic blacks. The outcome variable is the binarized annual healthcare expenditure (low cost and high cost), obtained by thresholding the expenditure at its national median (USD 1,272). Representative features considered for the classification problem include age, marital status, education, military status, self-reported health conditions, self-reported physical and cognitive limitations, employment status, poverty category, and insurance coverage status. The threshold t used for obtaining the TSP metric is 0.5 since the prevalence of outcomes in this data is equally balanced at 0.5. The disparity in

Table 1: MEPS outcome disparities between various groups in source (low-income) and target (high-income) populations. The disparity is given by $\Pr(Y = 1|G = k) - \Pr(Y = 1|G = l)$ with probabilities expressed as percentages.

Groups k and l	Source	Target
white males, black males	14.6	17.4
white males, white females	-12.8	-18.0
white males, black females	3.0	-3.4
black males, white females	-27.4	-35.4
black males, black females	-11.6	-20.9
white females, black females	15.8	14.5

prevalence of high cost outcomes among various gender–race intersections is provided in Table 1. The largest disparity is between black males and white females, whereas the smallest is between white males and black females.

We compare PCCS and TFCS with the baseline approaches in Table 2. The AUC on the target population for all methods are similar except for TFCS with $\lambda = 100$, as will be explained later.

The methods do show differences however in score disparity in the target. Covariate shift without any fairness adjustments happens to significantly reduce the fairness losses, and Kamiran-Calders yields a similar reduction. They are further reduced with PCCS, which combines elements of covariate shift and parity-inducing reweighting. TFCS with $\lambda = 100$ achieves by far the lowest score parity losses, at the cost of a lower AUC. The large setting for λ is intended to show the parity levels that can be achieved.

Figure 1 shows the variation of fairness and accuracy metrics with respect to δ for PCCS. While the MSP and TSP curves are somewhat variable, there is a slight downward trend in disparities as δ decreases toward zero and tightens the prevalence constraints (6). The AUC is nearly constant. The behavior of TFCS with changing λ is illustrated in Figure 2. Recall that λ weights the fairness component of the combined loss (12). When λ is small, all metrics are closer to the values obtained with methods that do not account for fairness (unadapted, covariate shift), as expected. As λ increases past 1, the score parity losses decrease dramatically while the AUC undergoes a more modest reduction.

Although we do not observe a boost in AUC from covariate shift methods, our fairness methods that adjust for covariate shift improve the fairness metrics over Kamiran-Calders, which does not account for covariate shift.

4.2 Mobile Money Loan Approval in East Africa

We consider the expansion of mobile-money credit services into a new market of East Africa. We use data from the *original* market to train a model that is deployed in the *new* market. Age (thresholded at 35 years) and gender are protected attributes. The prediction task is to identify who will repay a loan; the features are described in [22] and include airtime usage and mobile money volumes sent and received over a 6-month period. The threshold for approving a loan is $t = 0.75$ since banks will only issue a loan if they are confident a user will repay.

Table 4 illustrates that transfer learning significantly improves score parity in the target population, with covariate shift and even

Table 2: Results on MEPS dataset. Source: low-income population, Target: high-income population. Disparity measures are shown as percentages. All metrics are reported for target.

	AUC	MSP loss	TSP loss	Max Δ MSP	Max Δ TSP
Native	0.800	29.0	42.4	19.4	27.7
Unadapted	0.799	27.0	43.3	18.9	30.2
Covariate Shift	0.786	17.6	30.0	12.3	21.1
Kamiran Calders	0.799	16.5	26.2	11.5	18.5
PCCS ($\delta = 0.05$)	0.788	14.2	20.3	9.7	14.3
TFCS ($\lambda = 100$)	0.708	2.2	5.1	1.4	3.4

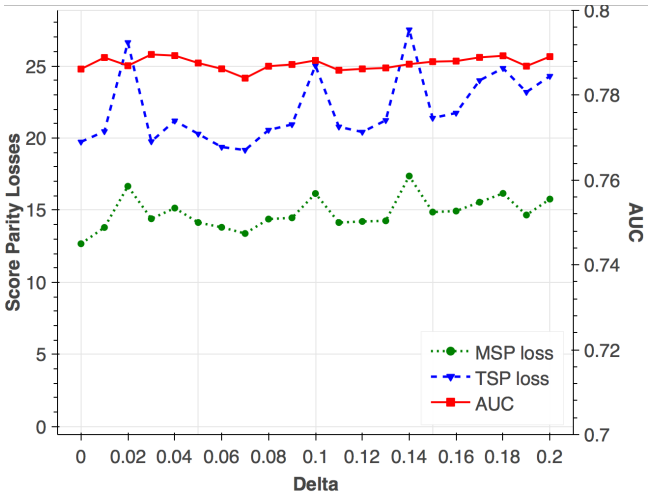

Figure 1: Variation of score parities and AUC with the PCCS trade-off parameter δ for the MEPS dataset.

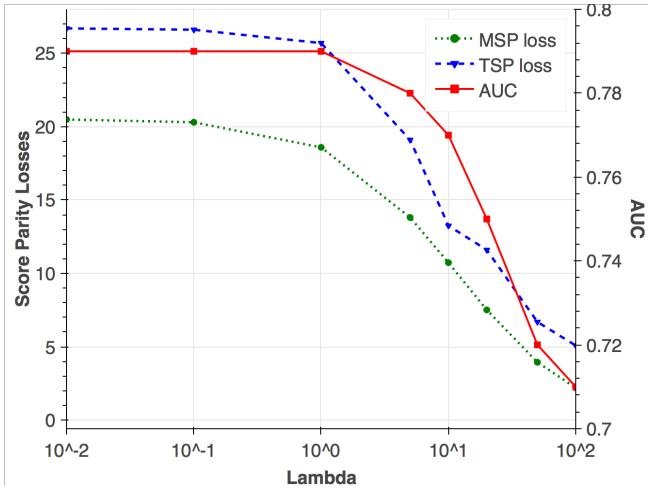

Figure 2: Variation of score parities and AUC with the TFCS trade-off parameter λ for the MEPS dataset.

in the unadapted case. We note that this is not because the source has less dataset bias; in fact, in Table 3, we see that the source dataset

Table 3: Mobile Money outcome disparities between various groups in source (original market) and target (new market) populations. The disparity is given by $\Pr(Y=1|G=k)-\Pr(Y=1|G=l)$ with probabilities expressed as percentages.

Groups k and l	Source	Target
female 35+, female under 35	7.0	6.2
female 35+, male 35+	3.5	1.4
female 35+, male under 35	11.1	9.2
female under 35, male 35+	-3.6	-4.8
female under 35, male under 35	4.1	3.0
male 35+, male under 35	7.7	7.9

Table 4: Results on Mobile Money dataset. Source: Country 1, Target: Country 2. Disparity measures are shown as percentages. All metrics are reported for target.

	AUC	MSP loss	TSP loss	Max Δ MSP	Max Δ TSP
Native	0.658	7.6	23.0	4.9	14.5
Unadapted	0.642	4.2	12.6	2.5	8.1
Covariate Shift	0.639	3.8	11.1	2.7	7.2
Kamiran Calders	0.640	4.0	12.5	2.5	7.7
PCCS ($\delta = .05$)	0.630	2.4	5.7	1.5	3.3
TFCS ($\lambda = .01$)	0.639	2.6	6.8	1.8	4.4

has larger disparities in loan approvals over the four demographic groups. PCCS and TFCS further reduce the four fairness metrics with little to no change in AUC for the target population. PCCS and TFCS outperform Kamiran-Calders, which only yields results on par with covariate shift. For this application, we found that increasing λ did not improve the score disparities for TFCS. Thus Table 4 shows results for $\lambda = 0.01$ which maintains the AUC.

Figure 3 shows the results for PCCS as we vary δ to trade off improving AUC against reducing the fairness losses. For $\delta \geq 0.08$, the weighted prevalence constraints (6) are loose enough to allow the covariate shift solution $w = w_{CS}$ and the metrics converge accordingly. The trade-off between AUC and score parity is seen for smaller δ.

5 CONCLUSION

This paper has discussed methods that address jointly the problem of covariate shift between source and target populations and the need to ensure fairness in a predictor's outputs toward protected groups. We have focused specifically on mean score parity and thresholded score parity measures of group fairness. Both of the proposed methods, prevalence-constrained covariate shift (PCCS) and target-fair covariate shift (TFCS), are based on sample reweighting and thus fit well with existing domain adaptation techniques and a variety of classification algorithms. Together they can accommodate the important practical limitation of having protected group information only in the source or target domain. Tested on two datasets, PCCS and TFCS show reductions in score disparity compared to baselines with little change in AUC. The MEPS dataset and mobile money credit dataset are new to the algorithmic fairness literature and, we believe, are more reflective of real risk assessment applications than some prior benchmarks.

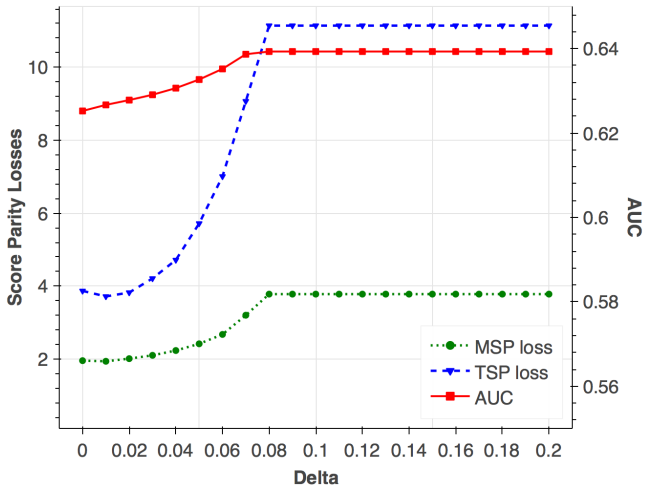

Figure 3: Variation of score parities and AUC with the PCCS trade-off parameter δ for the Mobile Money dataset.

ACKNOWLEDGMENTS

We acknowledge the helpful comments and feedback of Aldo Pareja. This work was conducted under the auspices of the IBM Science for Social Good initiative. This research in part was sponsored by the U.S. Army Research Lab and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *Proc. International Conference on Machine Learning (ICML)*. Stockholm, Sweden.
[2] Atılım Güneş Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. 2018. Automatic Differentiation in Machine Learning: A Survey. *Journal of Machine Learning Research* 18, 2 (2018), 1–43.
[3] Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative Learning Under Covariate Shift. *Journal of Machine Learning Research* 10 (Sept. 2009), 2137–2155.
[4] Kim M Blankenship. 1993. Bringing gender and race in: US employment discrimination policy. *Gender & Society* 7, 2 (1993), 204–226.
[5] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2018. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. *arXiv preprint arXiv:1811.11154* (2018).
[6] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163.
[7] Tamara Cook and Claudia McKay. 2015. How M-Shwari Works: The Story So Far. *Access to Finance Forum* 10 (April 2015).
[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax, NS, Canada, 797–806.
[9] Agency for Healthcare Research and Quality. 2018. Medical Expenditure Panel Survey (MEPS). http://www.ahrq.gov/research/data/meps/index.html.

[10] Jill Gaulding. 1995. Race Sex and Genetic Discrimination in Insurance: What's Fair. *Cornell Law Review* 80, 6 (Sept. 1995), 1646–1694.

[11] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. 2012. Geodesic Flow Kernel for Unsupervised Domain Adaptation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA, 2066–2073.

[12] Arthur Gretton, Alexander J Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M Borgwardt, and Bernhard Schölkopf. 2009. Covariate Shift by Kernel Mean Matching. (2009). https://doi.org/10.7551/mitpress/9780262170055.003.0008

[13] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy Fairness. arXiv:1806.11212.

[14] Samantha Kahn. 2015. The End of Gender Rating: Women's Insurance Under the ACA. https://publicpolicy.wharton.upenn.edu/live/news/819-the-end-of-gender-rating-womens-insurance-under.

[15] Faisal Kamiran and Toon Calders. 2012. Data Preprocessing Techniques for Classification Without Discrimination. *Knowledge and Information Systems* 33, 1 (Oct. 2012), 1–33.

[16] Pang Wei Koh and Percy Liang. 2017. Understanding Black-Box Predictions via Influence Functions. In *Proc. International Conference on Machine Learning*. Sydney, NSW, Australia, 1885–1894.

[17] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In *Proc. Web Conference*. Lyon, France, 853–862.

[18] Abhinav Maurya. 2018. IEEE Big Data 2017 Panel Discussion on Bias and Transparency. *AI Matters* 4, 2 (July 2018), 13–20.

[19] Jing Qin. 1998. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* 85, 3 (1998), 619–630.

[20] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (Eds.). 2009. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, USA.

[21] Hartley Rogers. 1999. *Multivariable Calculus with Vectors*. Prentice-Hall, Upper Saddle River, NJ, USA.

[22] Skyler Speakman, Srihari Sridharan, and Isaac Markus. 2018. Three Population Covariate Shift for Mobile Phone-Based Credit Scoring. In *Proc. ACM Conference on Computing and Sustainable Societies*. Menlo Park, CA, USA, 20.

[23] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. 2007. Covariate Shift Adaptation by Importance Weighted Cross Validation. *Journal of Machine Learning Research* 8 (May 2007), 985–1005.

[24] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. 2007. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems*. Vancouver, BC, Canada, 1433–1440.

[25] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. 2012. *Density Ratio Estimation in Machine Learning* (1st ed.). Cambridge University Press, New York, NY, USA.

[26] Xuezhi Wang and Jeff Schneider. 2014. Flexible Transfer Learning under Support and Model Shift. In *Advances in Neural Information Processing Systems 27*. Montreal, QC, Canada, 1898–1906.

[27] Sidney D. Watson. 2012. Section 1557 of the Affordable Care Act: Civil Rights, Health Reform, Race, and Equity. *Howard Law Journal* 55, 3 (Spring 2012), 855–886.

[28] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2015. Health Insurance Market Risk Assessment: Covariate Shift and k-Anonymity. In *Proc. SIAM International Conference on Data Mining*. Vancouver, BC, Canada, 226–234.

A GRADIENT DERIVATIONS FOR TFCS

The following derivation is similar to and takes inspiration from the theory of influence functions [16], which describe the effect of individual training points on model parameters. Here we consider the effect of re-weighting all training points at once.

First define $\mathcal{L}_c$ to be the re-weighted classification loss in (12),

$$\mathcal{L}_c(\hat{\theta}) = \frac{1}{n}\sum_{i\in\mathcal{S}} w_{CS}(x_i)\mathcal{L}\big(s(x_i;\hat{\theta}), y_i\big).$$

The derivative of $\mathcal{L}_c$ with respect to each w_i is given by the chain rule as

$$\frac{\partial\mathcal{L}_c}{\partial w_i} = \Big(\nabla_{\hat{\theta}}\mathcal{L}_c\Big)^T \frac{\partial\hat{\theta}}{\partial w_i}, \tag{13}$$

and similarly for $\mathcal{L}_f$. The second factor in (13) is the vector of partial derivatives $\partial\hat{\theta}_j/\partial w_i$ for all j. Hence for the combined loss $\mathcal{L}_t = \mathcal{L}_c + \lambda\mathcal{L}_f$,

$$\frac{\partial\mathcal{L}_t}{\partial w_i} = \Big(\nabla_{\hat{\theta}}\mathcal{L}_c + \lambda\nabla_{\hat{\theta}}\mathcal{L}_f\Big)^T \frac{\partial\hat{\theta}}{\partial w_i}. \tag{14}$$

To derive an expression for $\partial\hat{\theta}/\partial w_i$, we use the fact that if $\hat{\theta}$ is a minimizer in (10), then it must satisfy the first-order optimality conditions

$$\sum_{i\in\mathcal{S}} w_i \frac{\partial\mathcal{L}\big(s(x_i;\hat{\theta}), y_i\big)}{\partial\hat{\theta}_j} = 0 \quad \forall j.$$

For fixed $\{(x_i, y_i), i \in \mathcal{S}\}$, these conditions give a set of implicit equations for $\hat{\theta}$ in terms of $\{w_i\}$. We may obtain an *explicit* expression for the derivative $\partial\hat{\theta}/\partial w_i$ using the method of eliminating differentials [21, Ch. 11] as follows:

$$dw_i \frac{\partial\mathcal{L}\big(s(x_i;\hat{\theta}), y_i\big)}{\partial\hat{\theta}_j} + \sum_{i'\in\mathcal{S}} w_{i'} \sum_k \frac{\partial^2\mathcal{L}\big(s(x_{i'};\hat{\theta}), y_{i'}\big)}{\partial\hat{\theta}_j\partial\hat{\theta}_k} d\hat{\theta}_k = 0.$$

Rewriting in matrix-vector notation,

$$\nabla_{\hat{\theta}}\mathcal{L}\big(s(x_i;\hat{\theta}), y_i\big) + H_{\hat{\theta}}\frac{\partial\hat{\theta}}{\partial w_i} = 0,$$

where we have defined

$$H_{\hat{\theta}} = \sum_{i\in\mathcal{S}} w_i \nabla^2_{\hat{\theta}}\mathcal{L}\big(s(x_i;\hat{\theta}), y_i\big). \tag{15}$$

Hence

$$\frac{\partial\hat{\theta}}{\partial w_i} = -H^{-1}_{\hat{\theta}}\nabla_{\hat{\theta}}\mathcal{L}\big(s(x_i;\hat{\theta}), y_i\big). \tag{16}$$

For the case of binary classification with log loss $\mathcal{L}$ (aka cross-entropy) and logistic regression, we have

$$\mathcal{L}(s, y) = -y\log(s) - (1-y)\log(1-s),$$
$$s(x;\hat{\theta}) = \sigma(\hat{\theta}^T x) = \frac{1}{1+e^{-\hat{\theta}^T x}}.$$

where $\sigma(t)$ denotes the sigmoid function $\frac{1}{1+e^{-t}}$. Then

$$\nabla_{\hat{\theta}} s = \frac{e^{-\hat{\theta}^T x}}{\Big(1+e^{-\hat{\theta}^T x}\Big)^2}\cdot x = s(1-s)x, \tag{17}$$

$$\nabla_{\hat{\theta}}\mathcal{L} = (-y(1-s) + (1-y)s)x. \tag{18}$$

Using (18) and letting $s_i = s(x_i;\hat{\theta})$, the gradient of the classification loss is therefore

$$\nabla_{\hat{\theta}}\mathcal{L}_c = \frac{1}{n}\sum_{i\in\mathcal{S}} w_{CS}(x_i)[-y_i(1-s_i) + (1-y_i)s_i]x_i. \tag{19}$$

Likewise using (17), the gradient of the fairness loss is

$$\nabla_\theta\mathcal{L}_F = 2\sum_{k<l}\Bigg[\Bigg(\frac{1}{|\mathcal{T}_k|}\sum_{i\in\mathcal{T}_k} s_i - \frac{1}{|\mathcal{T}_l|}\sum_{i\in\mathcal{T}_l} s_i\Bigg)\times \Bigg(\frac{1}{|\mathcal{T}_k|}\sum_{i\in\mathcal{T}_k} s_i\big(1-s_i\big)x_i - \frac{1}{|\mathcal{T}_l|}\sum_{i\in\mathcal{T}_l} s_i\big(1-s_i\big)x_i\Bigg)\Bigg]. \tag{20}$$

For the Hessian we find

$$\nabla^2_{\hat{\theta}}\mathcal{L} = x\big(\nabla_{\hat{\theta}}s\big)^T = s(1-s)xx^T, \tag{21}$$

which is needed in (15).

To close this section, we justify the earlier statement that $w = w_{CS}$ is a stationary point in the case $\lambda = 0$. Since $\hat{\theta}$ is a minimizer in (10), it satisfies

$$\frac{1}{n} \sum_{i \in \mathcal{S}} w_i \nabla_\theta \mathcal{L}\big(s(x_i; \hat{\theta}), y_i\big) = 0.$$

The left-hand side coincides with $\nabla_{\hat{\theta}} \mathcal{L}_c$ when $w = w_{CS}$ and therefore $\nabla_{\hat{\theta}} \mathcal{L}_c = 0$. Combining this with $\lambda = 0$ in (14) implies that $\nabla_w \mathcal{L}_t = 0$ at $w = w_{CS}$.