Fairness in Semi-Supervised Learning: Unlabeled Data Help to Reduce Discrimination

Tao Zhang, Tianqing Zhu, Jing Li, Mengde Han,
Wanlei Zhou, *Senior Member, IEEE*, and Philip S. Yu, *Fellow, IEEE*

Abstract—A growing specter in the rise of machine learning is whether the decisions made by machine learning models are fair. While research is already underway to formalize a machine-learning concept of fairness and to design frameworks for building fair models with sacrifice in accuracy, most are geared toward either supervised or unsupervised learning. Yet two observations inspired us to wonder whether semi-supervised learning might be useful to solve discrimination problems. First, previous study showed that increasing the size of the training set may lead to a better trade-off between fairness and accuracy. Second, the most powerful models today require an enormous of data to train which, in practical terms, is likely possible from a combination of labeled and unlabeled data. Hence, in this paper, we present a framework of fair semi-supervised learning in the pre-processing phase, including pseudo labeling to predict labels for unlabeled data, a re-sampling method to obtain multiple fair datasets and lastly, ensemble learning to improve accuracy and decrease discrimination. A theoretical decomposition analysis of bias, variance and noise highlights the different sources of discrimination and the impact they have on fairness in semi-supervised learning. A set of experiments on real-world and synthetic datasets show that our method is able to use unlabeled data to achieve a better trade-off between accuracy and discrimination.

Index Terms—Fairness, discrimination, machine learning, semi-supervised learning

◆

1 INTRODUCTION

Machine learning is now in wide use as a decision-making tool in many areas, such as job employment, risk assessment, loan approvals and many other basic precursors to equity. However, the popularity of machine learning has raised concerns about whether the decisions algorithms make are fair to all individuals. For example, Chouldechova found evidence of racial bias in recidivism prediction tool where black defendants are more likely to be assessed with high risk than white defendants [1]. Obermeyer *et al.* found prejudice in health care systems where black patients assigned the same level of risk by the algorithm are sicker than white patients [2]. These findings show that unfair machine learning algorithms will affect legal justices, health care, and other aspects of human beings.

As we move forward in a world of machine-assisted predictions for human-beings, the fairness of machine learning has become a very cardinal issue. In the future, our ability to design machine learning algorithms that treat all groups equally may be one of the most influential factors in who will be the haves and who will be the have-nots. As the influence and scope of these risk assessments increase, academics, policymakers, and journalists have raised concerns that the statistical models from which they are derived might inadvertently encode human biases

Over the past few years, much research has been devoted to designing fairness metrics, such as statistical fairness [1], [3], [4], [5], individual fairness [6], [7], [8] and causal fairness [9], [10]. These approaches and algorithms can be roughly divided into three categories: pre-processing methods, in-processing methods and post-processing methods. Pre-processing methods adjust data distribution [3], [11] or learn new fair representations [12], [13], [14], to relieve some of the tension between accuracy and fairness. In-processing methods add constraints or regularizers to restrict the correlation between labels and sensitive/protected attributes, i.e., traits that can be targets for discrimination [4], [15], [16]. Post-process methods calibrate training results [5]. These studies mainly focus on addressing the two most crucial fundamental issues in machine learning fairness: how to formalize the concept of fairness in the context of machine learning tasks, and how to design effective algorithms to achieve an ideal compromise between accuracy and fairness.

However, almost all methods achieving fairness are mostly for either supervised learning or unsupervised learning, and fair semi-supervised learning (SSL) has rarely been considered. Realistically though, training data is often a combination of labeled and unlabeled samples, so a semi-supervised solution has high practical value. Also, since "ideal" is a lofty goal, the trade-off between accuracy and fairness is still an ongoing pursuit. [17] showed that increasing the amount of training data is likely to produce a better trade-off between accuracy and fairness. This insight inspired us to wonder whether using unlabeled data to augment the training set might give us a kind of control value with which to

---

- *T. Zhang, T. Zhu, M. Han, and W. Zhou are with Centre for Cyber Security and Privacy, School of Computer Science, University of Technology, Sydney, Ultimo, NSW 2007, Australia. E-mail: {Tao.Zhang-3, Mengde.Han}@student.uts.edu.au, {Tianqing.Zhu, Wanlei.Zhou}@uts.edu.au.*
- *J. Li is with the Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia. E-mail: jing.li-20@student.uts.edu.au.*
- *P.S. Yu is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607 USA. E-mail: psyu@uic.edu.*

Manuscript received 23 Mar. 2020; revised 27 May 2020; accepted 11 June 2020. Date of publication 15 June 2020; date of current version 7 Mar. 2022.
(Corresponding author: Tianqing Zhu.)
Recommended for acceptance by W. Zhang.
Digital Object Identifier no. 10.1109/TKDE.2020.3002567

balance fairness and accuracy. Unlabeled data is abundant and, if it could be used as training data, we could adjust the size of the training set as required to meet accuracy versus fairness thresholds. We may even be able to avoid the need to make a compromise between fairness and accuracy entirely. This leaves fair semi-supervised learning with two challenges: 1) How to make use of unlabeled data to achieve a better trade-off between accuracy and fairness; and 2) How to alleviate the impact of noise, which is common to semi-supervised learning.

To tackle these challenges, we propose a framework to achieve fair SSL in the pre-processing phase. The solution to the trade-off challenge is to use unlabeled data to reduce representation discrimination. (Representation discrimination is due to certain parts of the input space under-represented.) Therefore, the first two steps in our framework are pseudo labeling and re-sampling. The first step is to use pseudo labeling as a SSL method to predict labels for unlabeled data. The second step involves dividing the dataset into groups based on the protected attribute and the label, and then obtain fair datasets by re-sampling the same number of data points in each group. When unlabeled data is used as training data, it is likely to obtain more under-represented data points from unlabeled data to reduce representation discrimination, and thus to make little compromise between fairness and accuracy. The issue of noise induced by (incorrectly) predicting labels for unlabeled data is addressed by the third step in the framework: ensemble learning. Predicting unlabeled data will induce some noise in the labels of unlabeled data. Ensemble learning helps to reduce label noise and the variance of the training model, and to produce more accurate final predictions.

In summary, the contributions of this paper are listed as below.

- First, we use unlabeled data to reduce representation discrimination, and thus achieve a better trade-off between accuracy and discrimination.
- Second, we propose a fairness-enhanced sampling (FS) framework that combines pseudo labeling, re-sampling and ensemble learning for fair SSL in the pre-processing phase.
- Third, we theoretically analyze the sources of discrimination in SSL via bias, variance and noise decomposition, and conduct experiments with both real and synthetic data to validate the effectiveness of our proposed FS framework.

The rest of this paper is organized as follows. The background is presented in Section 2, and the proposed FS framework is given in Section 3. Section 4 presents the discrimination analysis, and the experiments are set out in Section 5. The related work appears in Section 6, with the conclusion in Section 7.

2 BACKGROUND

2.1 Notations

For simplicity, let $\mathcal{D}_l = \{X, A, Y_l\}$ be a dataset with N_1 data points, where $X = (X_1, X_2, \ldots, X_d)$ denotes d unprotected attributes; A denotes protected attributes, e.g., gender or race; and $Y_l \in \{0, 1\}$ is the label for the task. Let $\mathcal{D}_u = \{X, A, Y_u\}$ be an unlabeled dataset with N_2 data points and $Y_u \in \{0, 1\}$ be the predicted labeled for the unlabeled dataset. For ease, assume the protected attribute is binary valued. For example, if the protected attribute is race, the value might be either 'white' ($A = 0$) or 'black' ($A = 1$).

Our objective is to learn a mapping $f(\cdot)$ over a discriminatory dataset $\mathcal{D}_l$ and $\mathcal{D}_u$, in which the classification result is independent of protected attributes. Performance is measured by both accuracy and the level of discrimination in the results. The ideal classifier should have a high accuracy without discrimination.

2.2 Fairness Metrics

Fairness is often evaluated with respect to protected/unprotected groups of individuals defined by attributes, such as gender or age. Here, we have opted for demographic parity as the fairness metrics in this paper.

Definition 1 (Demographic parity). *[3] Demographic parity requires that the probability of a classifier's prediction be independent of any sensitive attributes, where the probability of the predicted positive labels in group $a \in \mathcal{A}$ is defined as follows:*

$$\gamma_1(\hat{Y}) = Pr(\hat{Y} = 1 | A = 1) \tag{1}$$

$$\gamma_0(\hat{Y}) = Pr(\hat{Y} = 1 | A = 0). \tag{2}$$

Definition 2 (Discrimination level). *The discrimination level γ in terms of demographic parity can be evaluated by the difference between groups,*

$$\Gamma(\hat{Y}) = |\gamma_0(\hat{Y}) - \gamma_1(\hat{Y})|. \tag{3}$$

2.3 Discrimination Sources

Discrimination can exist in every stage of machine learning. Roughly, discrimination sources can be divided into two lines: data discrimination and model discrimination [18]. Our proposed FS method is able to reduce the representation discrimination in the data.

2.3.1 Data Discrimination

Data discrimination includes historical discrimination, representation discrimination, measurement discrimination. Historical discrimination occurs when there is a discrepancy between the world itself and the values or goals in the model to be encoded and propagated. It can stem from cultural stereotypes among people, such as social class, race, nationality, gender. Representation discrimination occurs when the data used to train the algorithm does not accurately represent the problem space. As a consequence, the model generalizes to fit the majority groups much than minority groups. Measurement discrimination comes from the way we choose, utilize, and measure specific features. The selected set of features and labels may miss important factors, or bring in group or input-related noise that causes different performance.

2.3.2 Model Discrimination

Model discrimination includes aggregation discrimination, evaluation discrimination, deployment discrimination. Aggregation discrimination can arise during model

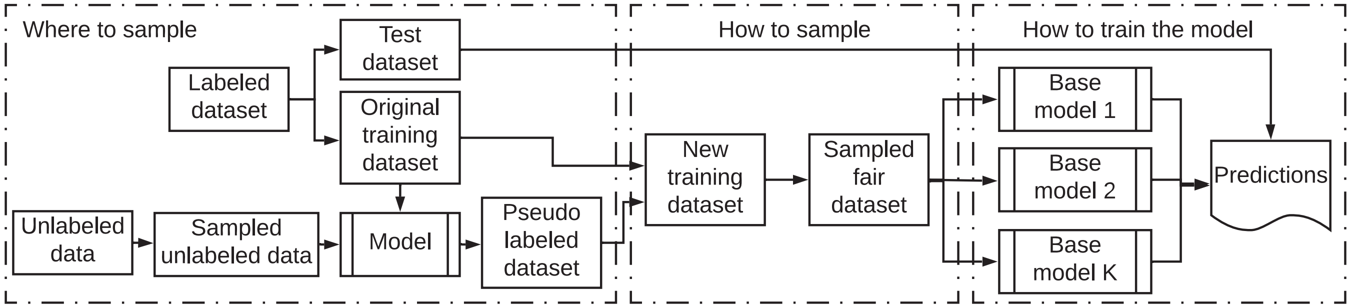


Fig. 1. The three phases of the fairness-enhanced sampling framework: 1) where to sample, 2) how to sample and 3) how to train the model. Step 1 is to generate a new training dataset which consists of the original dataset and the pseudo labeled dataset. Step 2 is to construct multiple fair datasets through re-sampling. Step 3 is to train a model with each of the fair datasets through ensemble learning to produce the final predictions.

construction when different populations are improperly grouped together. In many applications, the groups of interest are heterogeneous, so a single model is unlikely to fit all subgroups. Evaluation discrimination occurs during model iteration and evaluation. This can happen when a test or external benchmark unequally represents each group in the population. Evaluation discrimination may also occur due to the use of performance metrics that are not appropriate for the way the model is used. Deployment discrimination occurs after the model is deployed when the system is used or interpreted in an inappropriate way.

2.4 Bias, Variance and Noise

Following [17], our analysis of discrimination is based on bias, variance and noise decomposition. First, we present the definition of main prediction. The main prediction for a loss function L and set of training sets D is defined as, $y_m(x,a) = argmin_{y'} \mathbb{E}_D[L(Y,y')|X=x, A=a]$, where Y is the true value; y' is the predicted label with the minimum average loss relative to all the predictions. The expectation is taken with respect to the training sets in D.

Definition 3. *(Bias,variance and noise) Following [19], the bias B, variance V and noise N at a point (x,a) with a model f are defined as,*

$$B(f,x,a) = L(y^*(x,a), y_m(x,a)) \tag{4}$$

$$V(f,x,a) = \mathbb{E}_D[L(y_m(x,a), \hat{y}_D(x,a)] \tag{5}$$

$$N(f,x,a) = \mathbb{E}_Y[L(y^*(x,a), Y)], \tag{6}$$

where y^* is the optimal prediction that achieves the smallest expected error. Bias is the loss between the main prediction and the optimal prediction. Variance is the average loss incurred by predictions relative to the main prediction from different datasets D. Noise is the unavoidable component of the loss, which is independent of the learning model.

Bias, variance and noise decomposition are appropriate tools for analyzing discrimination because loss function relates to the misclassification rate. For example, when using zero-one loss function, the misclassification rate is denoted as

$$\begin{aligned}\mathbb{E}[L(y,\hat{y})] &= \mathbb{E}[\hat{y} \neq y|a=0] + \mathbb{E}[\hat{y} \neq y|a=1] \\ &= \mathbb{E}[\hat{y}=1|y=0,a=0] + \mathbb{E}[\hat{y}=0|y=1,a=1],\end{aligned} \tag{7}$$

where $\hat{y}$ is the predicted label of a classifier. Note that loss function can be decomposited into false positive rate and false negative rate. And, once false positive rate and false negative rate are obtained, true positive rate and true negative rate can be obtained. As such, many fairness metrics, such as demographic parity and equal opportunity, can be explained by bias, variance and noise decomposition.

3 THE PROPOSED METHOD

3.1 Overview of the Fairness-Enhanced Sampling Framework

Fig. 1 shows the general description of the fairness-enhanced sampling framework in the pre-precessing phase. The framework consists of three steps: 1) pseudo labeling, 2) resampling and 3) fair ensemble learning. The first step is to predict labels for unlabeled data as more data points in the protected group are likely to be found in unlabeled data. The second step is to construct new datasets that is able to represent all groups equally when the datasets are used for training. In this way, representation discrimination can be removed from training datasets. The third step is to train multiple base models based on multiple fair datasets and final predicted results are obtained from multiple base models. Ensemble learning is able to reduce the label noise that is induced via pseudo labeling, and the model variance. Each of these steps is discussed in more detail in the following.

3.2 Where to Sample

The goal of this step is to use a labeled dataset and part of an unlabeled dataset to construct a new training dataset, as shown in Fig. 1. Suppose we have a labeled dataset $\mathcal{D}_l$ and a large unlabeled dataset $\mathcal{D}_u$. First, we use the labeled dataset and part of the unlabeled dataset to generate a new training dataset. With a sample ratio of ρ, we take random samples from the unlabeled dataset $\mathcal{D}_u$ and form sampled unlabeled datasets $\mathcal{D}_{su}$. Then we use pseudo labeling to predict the labels for unlabeled data as if they were true labels. Pseudo labeling is a simple and efficient method to implement SSL [20]. The procedure, as shown in Algorithm 1, is as follows.

1) Set a split rate $s \in (0,1)$ and split the labeled dataset into training and test dataset, denoted as the original training dataset and test dataset. 2) Select a learning model and, train the model on the original training dataset to produce a trained model. 3) Use the trained model on $\mathcal{D}_{su}$ to predict the output (or pseudo label), and the pseudo labeled dataset is obtained. We do not know if these predictions are correct,

but we now have predicted labels, which is our goal in this step. 4) Concatenate the original training dataset and pseudo labeled dataset to form a new training dataset $\mathcal{D}_{new}$.

Algorithm 1. Pseudo Labeling

```
Input: Labled dataset $\mathcal{D}_l$, unlabeled dataset $\mathcal{D}_u$, split rate $s$,
       sample ratio $\rho$
Output: New training dataset $\mathcal{D}_{new}$
Split $\mathcal{D}_l$ into the training dataset and the test dataset;
Sample $\mathcal{D}_{su}$ from $\mathcal{D}_u$;
Select a learning model and train the model on the training
  dataset;
Obtain the trained model;
Use the trained model to predict labels for $D_{su}$;
Combine the original training dataset and the pseudo
  labeled dataset to create $D_{new}$;
```

Pseudo-labeling is an easy-to-implement and efficient semi-supervised learning method and, by the above method, can take advantage of unlabeled data to both: a) increase the size of the training set; and b) create more data samples representing minority groups to produce fairer training sets. Moreover, the learning model can be any models, such as logistic regression, neural networks, etc.

3.3 How to Sample

In this step, the goal is to sample multiple fair datasets from the new training datasets to ensure fair learning. The rationale for this method is that, since the classifier is trained on non-discriminatory data, its prediction may also be non-discriminatory [11]. For simplicity, this analysis covers a binary classification task with one protected attribute, and applies demographic parity as the fairness metric. Our method can certainly be applied to cases with multiple sensitive attributes, subjected to the fairness metrics.

Based on this setup, the dataset is divided into four groups according to the protected attribute and labeled-values: 1) Protected group with positive labels (G_{PP}), 2) Unprotected group with positive labels (G_{UP}), 3) Protected group with negative labels (G_{PN}), and 4) Unprotected group with negative labels (G_{UN}). These divided groups can be denoted as follows,

$$G_{PP} = \{X \in D | A = 1, Y = 1\} \tag{8}$$

$$G_{UP} = \{X \in D | A = 0, Y = 1\} \tag{9}$$

$$G_{PN} = \{X \in D | A = 1, Y = 0\} \tag{10}$$

$$G_{UN} = \{X \in D | A = 0, Y = 0\}, \tag{11}$$

where $Y = 1$ denotes the positive class and $Y = 0$ denotes the negative class. $A = 1$ denotes that the data point is in the protected group and $A = 0$ denotes that the data point is in the unprotected group. To ensure fair learning in the pre-processing phase, the number of data points in the training set for each group should be the same, otherwise the model will fall prey to data discrimination. In the case of discrimination, the size of each group is different. Our aim is to adjust the data points by sampling to reach the same size in each group.

Algorithm 2 describes the process of how to obtain multiple fair datasets, and the procedure is as follows: First, we compute the size of the groups G_{PP}, G_{UP}, G_{PN}, G_{UN}. The sample size is denoted as n_s, which means that the number of n_s data points will be sampled from each group. Here, there are two cases: 1) When $n_i \geq n_s$, n_s data points are sample randomly from the group G_i. 2) When $n_i < n_s$, n_s data points are oversampled from the group G_i. Then we can obtain the fair dataset $\mathcal{D}_{sf}$ which consists of the number of data points equally for each of the four groups. Repeating this procedure K times produces K fair datasets with some commonalities and some differences due to the random sampling, which is desirable for ensemble learning. The next step is to learn from these multiple fair datasets to achieve more accurate and less discriminatory results.

Algorithm 2. Fair Re-sampling

```
Input: New training datase $\mathcal{D}_{new}$, sensitive attribute $A$, sample
       times $K$, sample size $n_s$, sample ratio $\rho$
Output: Fair datasets $\mathcal{D}_{sf}$
Divide the dataset into four groups $G_{PP}$, $G_{PN}$, $G_{UP}$, $G_{UN}$
Calculate the size of all groups $n_i$
for $k \in K$ do
    if $n_i \geq n_s$ then
        Sample randomly the number of $n_s$ data points from
          the group $i$
    end
    if $n_i \leq n_s$ then
        Oversample the number of $n_s$ data points from the
          group $i$
    else
    end
    Obtain fair datasets $\mathcal{D}_{sf,i}$
end
Obtain multiple fair datasets $\mathcal{D}_{sf,1}, \mathcal{D}_{sf,2}, \ldots, \mathcal{D}_{sf,K}$
```

3.4 How to Train the Model

In this step, the goal is to achieve more accurate and less discriminatory training results on multiple fair datasets $\mathcal{D}_{sf}$. After obtaining multiple $\mathcal{D}_{sf}$, we choose a learning model to train multiple $\mathcal{D}_{sf}$ and apply ensemble learning to combine the learning results. Ensemble learning in machine learning exploits the independence between base models to improve the overall performance. In this case, we use Bagging [21] to combine the decisions from multiple base models learned on multiple fair datasets to improve the accuracy and decrease the discrimination.

Algorithm 3 describes the fair ensemble learning. With the new training dataset $\mathcal{D}_{new}$ from Algorithm 1 and fair datasets $\mathcal{D}_{sf,1}, \mathcal{D}_{sf,2}, \ldots, \mathcal{D}_{sf,K}$ from Algorithm 2, train each fair dataset on its own model $f_k(\mathcal{D}_{sf,k})$ in parallel. The final model will average the outputs based on the aggregation of predictions from all base models. The predictions obtained from most base models are predicted as final predictions, which is presented as,

$$f(\cdot) = argmax_{y \in \mathcal{Y}} \sum_{k=1}^{K} \mathbb{I}(y = f_k(D_{sf,k})), \tag{12}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and K is the ensemble size, i.e., the number of fair datasets.

Having some diversity across the datasets is crucial for ensemble learning. In our approach, the randomness of the fair datasets reflects in two places: 1) randomly sampling the unlabeled dataset $\mathcal{D}_u$, and subsequently, the pseudo labeled dataset process in Algorithm 1; and 2) randomly sampling n_s data points for all groups from $\mathcal{D}_{new}$ when constructing each fair dataset.

With ensemble learning, the discrimination level is determined by final predictions. We redefine the discrimination level in ensemble learning as $\gamma_{En} = |Pr(f(\cdot) = 1|A = 1) - Pr(f(\cdot) = 1|A = 0)|$. Overall, a combination of multiple base models helps to decrease discrimination resulting from variance and noise, and is able to give a more reliable prediction than a single model.

Algorithm 3. Fair Ensemble Learning

```
Input: Dataset, sample times $K$, sample size $n_s$, split rate $s$,
       sample ratio $\rho$
Output: Accuracy $Acc$, Discrimination $\gamma$
1: Execute Algorithm 1 to obtain the new training dataset $\mathcal{D}_{new}$;
2: for $k \in K$ do
3:     Execute Algorithm 2 to obtain the fair dataset $\mathcal{D}_{sf,k}$;
4:     Train the selected model on the fair dataset $\mathcal{D}_{sf,k}$ and
       obtain the base model $f_k(\cdot)$
5: end
6: Make predictions using the final model with ensemble size
   $K$ in Eq. (12);
```

3.5 Discussion

In reviewing the complete framework, there are several benefits to this approach, which are worth highlighting.

- Many semi-supervised learning methods can be used to predict labels for unlabeled data, such as graph-based learning and transductive support vector machines [22]. We choose pseudo labeling because it is a commonly used semi-supervised learning technique, which is efficient and easy to implement.
- The proposed FS framework only removes representation discrimination. However, it is likely that many types of discrimination exist in machine learning, such as historical discrimination, measurement discrimination. Other discrimination can be removed by in-processing or post-processing methods, based on our proposed FS framework.

4 DISCRIMINATION ANALYSIS

Following [17], we analyze the fairness of the predictive model via bias, variance, and noise decomposition. The source of discrimination can be decoupled as discrimination in bias $B_a(f)$, discrimination in variance $V_a(f)$ and discrimination in noise N_a. The expected discrimination level $\Gamma(f)$ of a classifier f learned from a set of training set D is defined as, $\bar{\Gamma}(f) = |\mathbb{E}_D[\Gamma_0(f) - \Gamma_1(f)]|$.

Lemma 1. *The discrimination with regard to group $a \in A$ is defined as,*

$$\gamma_a(f) = \bar{B}_a(f) + \bar{V}_a(f) + \bar{N}_a. \tag{13}$$

Given two groups, the discrimination level is denoted as,

$$\bar{\Gamma} = |(\bar{B}_0(f) - \bar{B}_1(f)) + (\bar{V}_0(f) - \bar{V}_1(f)) + (\bar{N}_0 - \bar{N}_1)|.$$

And, in more detail, the discrimination components of Eq.(13), i.e., bias, variance and noise are as follows:

$$\bar{B}_a(f) = \mathbb{E}_D[B(y_m, x, a)|A = a] \tag{14}$$

$$\bar{V}_a(f) = \mathbb{E}_D[c_v(x, a)V(y_m, x, a)|A = a] \tag{15}$$

$$\bar{N}_a = \mathbb{E}_D[c_n(x, a)L(y^*(x, a), Y)|A = a], \tag{16}$$

where $c_v(x, a)$ and $c_n(x, a)$ are parameters related to the loss function. For more details, see the proof in [17].

Lemma 2. *The discrimination learning curve $\bar{\Gamma}(f, n) := |\bar{\gamma}_0(f, n) - \bar{\gamma}_1(f, n)|$ is asymptotic and behaves as inverse power law curve, where n is the size of the training data [17].*

Theorem 1. *Unlabeled data is able to reduce discrimination with the proposed FS framework, if $(|\bar{V}_a(f)_{sl}| - |\bar{V}_a(f)_{ssl}|) - \bar{N}_{a,p} \geq 0$.*

Proof. To prove the above theorem, we shall prove that the discrimination level in SSL $\bar{\Gamma}_{ssl}$ is lower than the discrimination level in supervised learning $\bar{\Gamma}_{sl}$. In the following, we will analyze the discrimination in SSL in terms of *discrimination in bias* $\bar{B}_a(f)_{ssl}$, *discrimination in variance* $\bar{V}_a(f)_{ssl}$, *and discrimination in noise* $\bar{N}_{a,ssl}$.

Discrimination in Bias. Bias measures the fitting ability of the algorithm itself, and describe accuracy of the model. Hence, bias in discrimination $\bar{B}_a(f) = \mathbb{E}_D[B(y_m, x, a)|A = a]$ only depends on the model. When the same model is trained on the original training dataset and new training dataset, discrimination in bias is the same in supervised learning and SSL, which can be expressed as $|\bar{B}(f)_{sl}| - |\bar{B}(f)_{ssl}| = 0$.

Discrimination in Variance. Discrimination in variance $\bar{V}_a(f)$ can be reduced with extra unlabeled data in the training dataset. Lemma 2 states that the discrimination level $\bar{\Gamma}(f, n)$ decreases with the increasing size of training data n. In our proposed FS framework, unlabeled data is pseudo-labeled, and the new training dataset consists of the original training dataset and the pseudo labeled dataset. The size of the new training dataset can be guaranteed to be larger than the size of the original training by adjusting the sampling size. Also, using Bagging to combine all the base models to obtain the final predictions helps to construct the aggregate model with a lower variance, thus reducing the discrimination in variance $\bar{B}_a$. Hence, we conclude that $|\bar{V}_a(f)_{ssl}| - |\bar{V}_a(f)_{sl}| \leq 0$.

Discrimination in Noise. Unlabeled data introduces more discrimination in noise because pseudo labeling contains discrimination from the trained model. Thus, noisy labels from pseudo labeling in the unprotected group is more than that in the protected group. We divide the discrimination in noise in SSL into discrimination in noise in labeled data $\bar{N}_{a,l}$ and discrimination in noise in pseudo labeled data $\bar{N}_{a,p}$, which is expressed as,

$$\bar{N}_{a,ssl} = \bar{N}_{a,l} + \bar{N}_{a,p}. \tag{17}$$

Discrimination in noise in labeled data $\bar{N}_{a,l}$ is the same as the discrimination in noise in supervised learning $\bar{N}_{a,sl}$. Then we analyze the discrimination in noise due to pseudo labeled data $\bar{N}_{a,p}$, including four mislabeled cases,

$$\bar{N}_{y=0,a=0} = \mathbb{E}_{D_{un}}[\hat{y}_p^* = 1 | y = 0, a = 0] \tag{18}$$

$$\bar{N}_{y=0,a=1} = \mathbb{E}_{D_{un}}[\hat{y}_p^* = 1 | y = 0, a = 1] \tag{19}$$

$$\bar{N}_{y=1,a=0} = \mathbb{E}_{D_{un}}[\hat{y}_p^* = 0 | y = 1, a = 0] \tag{20}$$

$$\bar{N}_{y=1,a=1} = \mathbb{E}_{D_{un}}[\hat{y}_p^* = 0 | y = 1, a = 1], \tag{21}$$

where $\hat{y}_p^*$ is the optimal predicted label of unlabeled data via pseudo labeling. The noise in the protected group is $\bar{N}_{1,p} = \bar{N}_{y=0,a=1} + \bar{N}_{y=1,a=1}$ and the noise in the unprotected group is $\bar{N}_{0,p} = \bar{N}_{y=0,a=0} + \bar{N}_{y=1,a=0}$. The model contains discrimination because the model is trained on a dataset without any fairness guarantees, and thus the model will bring discrimination in pseudo labeling. In this way, discrimination in noise in pseudo labeled data $\bar{N}_{a,p}$ can be measured as,

$$\bar{N}_{a,p} = |\bar{N}_{1,p} - \bar{N}_{0,p}|. \tag{22}$$

To relieve the noise from pseudo labeling, we use Bagging—a robust model that is resilient to class label noise since the errors incurred by the noise can be compensated by the combined predictions of other learners.

Based on the analysis above, we conclude that when $|\bar{V}_a(\hat{Y})_{ssl} - \Delta\bar{V}_a(\hat{Y})_{sl}| - \bar{N}_{a,p} \geq 0$, unlabeled data is able to reduce discrimination with the proposed FS framework. Unlabeled data do not change discrimination in bias. However, they do reduce discrimination in variance, and they increase discrimination in noise, but bagging reduces discrimination both in variance and discrimination in noise. □

5 EXPERIMENT

In this section, we demonstrate our framework by performing experiments on real-world and synthetic datasets. The goal of our experiments is three folds. The first is to show how the framework makes use of unlabeled data to achieve a better trade-off between accuracy and discrimination. The second is to explore the impact of factors, such as ensemble times and sampling size, on the training results. And, third, we show the distinct difference in discrimination level when the model is tested with discrimination test dataset and fair test dataset.

5.1 Experiments on Real Data

The aim of real-world datasets is to assess the effectiveness of our method to achieve a better trade-off between accuracy and discrimination with unlabeled data. We also show the benefit of ensemble learning, the impact of the sampling size, and the comparison with other methods.

5.1.1 Experimental Setup

Dataset. The experiments involve three real-world datasets: the Health dataset,[1] the Bank dataset,[2] the Adult dataset.[3]

- The target of Health dataset is to predict whether people will spend any day in the hospital. In order to convert the problem into the binary classification task, we simply predict whether people will spend any day in the hospital or not. Here, 'Age' is the protected attribute and two groups are divided at ≥65 years. After data pre-processing, the dataset contains 10,000 records with 132 features.
- The Bank dataset contains a total of 31,208 records with 20 attributes and a binary label, which indicates whether the client has subscribed to a term deposit or not. Again, 'Age' is the protected attribute.
- The target of Adult dataset is to predict whether people's income is larger than 50K dollars or not, and we consider "Gender" as the protected attribute. After data pre-processing, the dataset contains 48,842 records with 18 features.

Parameters. The protected attribute is excluded from the prediction model during the training to ensure equity across groups. The protected attribute is only used to evaluate the discrimination measurement in the testing phrase. In the above of three real-world datasets, data are all labeled. First, we split the whole dataset randomly into two halves: one half is used as labeled dataset, and we remove the labels from the other half to served as the unlabeled dataset. In the labeled data, we set the split rate $s = 0.8$, which means 80 percent of the data are used for training and 20 percent of the data are used for testing. The sample size n_s equals the minimum size of four groups in three datasets.

The final result is an average of 50 results run in the new training datasets. For each run, we generate $K = 200$ fair datasets and construct with $K = 200$ base models to make the final predictions. We use 5-fold cross-validation on the original training dataset and test dataset.

Baseline. Given our method is a pre-processing method, we compare it to two other pre-processing methods and the method without any fairness process.

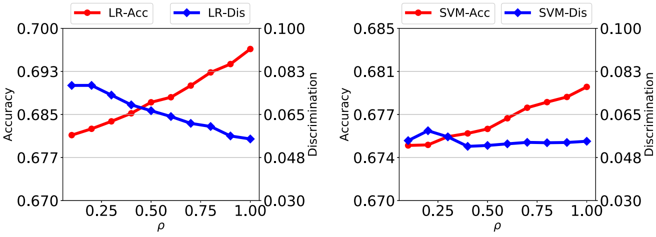
- Original (ORI): The original dataset is used for training without fairness guarantees.
- Uniform Sampling (US) [11]: The number of data points in each groups is equalized through oversampling and/ undersampling.
- Preferential Sampling (PS) [11]: The number of data points in each groups is equalized by taking samples near the borderline data points.

5.1.2 Trade-Off Between Accuracy and Discrimination

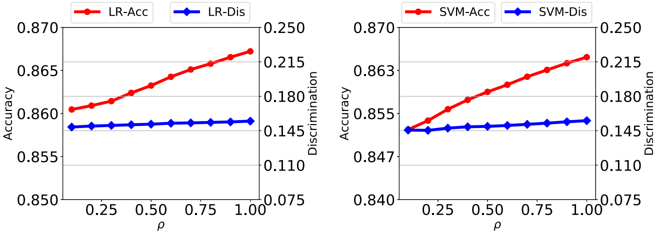
Fig. 2 shows the accuracy and discrimination level varies given different sample ratio ρ with logistic regression (LR)

---

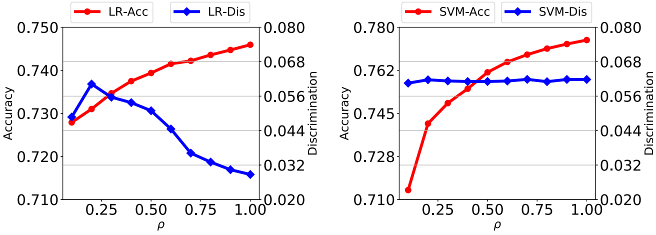
[1] https://foreverdata.org/1015/index.html
[2] https://archive.ics.uci.edu/ml/datasets/bank+marketing
[3] http://archive.ics.uci.edu/ml/datasets/Adult



(a) LR-Health (b) SVM-Health



(c) LR-Bank (d) SVM-Bank



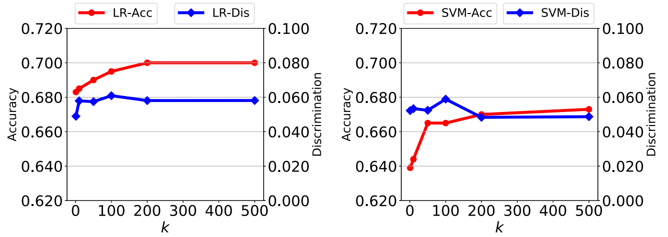
(e) LR-Adult (f) SVM-Adult

Fig. 2. The trade-off between accuracy (Red) and discrimination level (Blue). (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. The X-axis is the sample ratio ρ, which denotes that the percentage of ρ unlabeled data are sampled from the unlabeled dataset and then pseudo labeled for training.

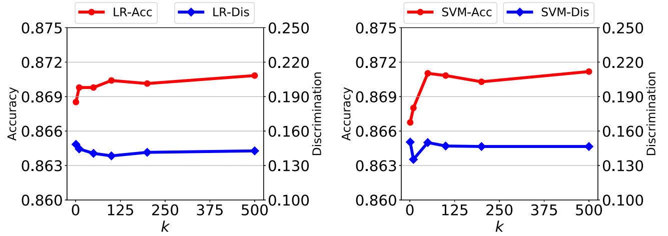
and support vector machine (SVM) on three datasets. As shown, accuracy generally increases with a growing size of unlabeled data. For example, LR has an accuracy of around 0.728 when $\rho = 0.1$ with the Adult dataset, which increases to 0.745 when $\rho = 1$. This indicates that the unlabeled data helps to improve the accuracy to some extent. Also, we note that accuracy relates to the training models and the choice of training models relates to the datasets. The discrimination level has different performances in different training models. For example, with the Adult dataset, the discrimination level initially increases and then steadily decreases till the end in LR. The discrimination level is steady and has a slight increase in SVM. This observation indicates that unlabeled data can help to reduce the discrimination for some models, like LR. Similar to accuracy, the discrimination level relates to the training models and our experiments show that LR is more friendly in discrimination than SVM. The choice of sample ratio depends on the quality of the dataset itself as well as the requirement of the learning task. Accuracy could be improved with unlabeled data, while discrimination level depends on the reduction of discrimination in variance and increase of discrimination in noise that unlabeled data could bring in the training.

5.1.3 The Impact of Ensemble Learning

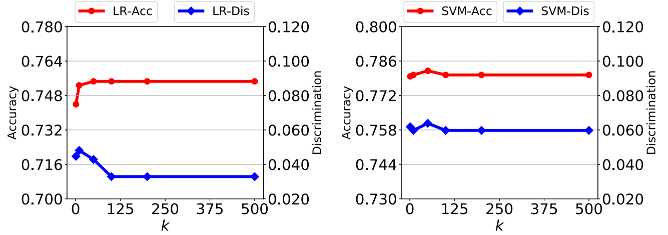
Fig. 3 shows the impact of ensemble learning on accuracy and discrimination level with LR and SVM on three datasets.



(a) LR-Health (b) SVM-Health



(c) LR-Bank (d) SVM-Bank



(e) LR-Adult (f) SVM-Adult

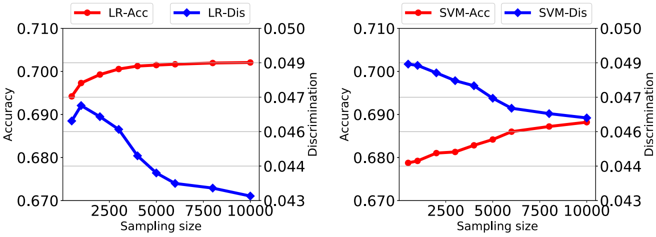
Fig. 3. The impact of ensemble learning on the accuracy (Red) and discrimination level (Blue) on (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. Initially, there is not obvious link between accuracy and discrimination level. However, as the ensemble size grows, the accuracy and discrimination level begin to converge. Each point is an average of 50 times.

In ensemble learning, we sample percentage of $\rho = 1$ unlabeled data from the unlabeled dataset, and generate the new training dataset. With LR, the accuracy typically increases then steadies till the end, whereas, with SVM accuracy fluctuates before steadying at some lower, equal or higher rate. This is because the errors in variance and noise reduce as the ensemble size increases.

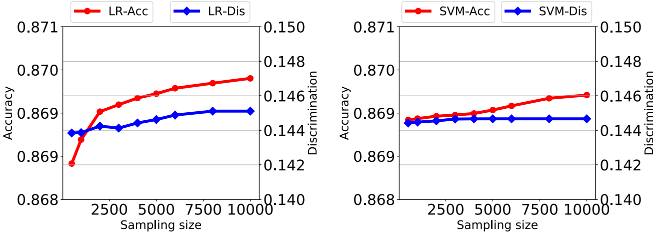
In terms of discrimination levels, both methods show fluctuations at first before stabilizing on all three datasets. The changes in discrimination levels have no obvious correlations to accuracy prior to convergence. This is reasonable because training results having the same accuracy does not mean the same discrimination level. Also, without a sufficient ensemble size, training on fair datasets will introduce some variance and noise to the final result. Overall, an ample ensemble size helps to improve accuracy and decrease discrimination. The appropriate ensemble size is $K = 200$ or so. This is because accuracy increases and discrimination fluctuates before $K = 200$, and broadly accuracy and discrimination become steady after $K = 200$ for three datasets.

5.1.4 The Impact of Sample Size

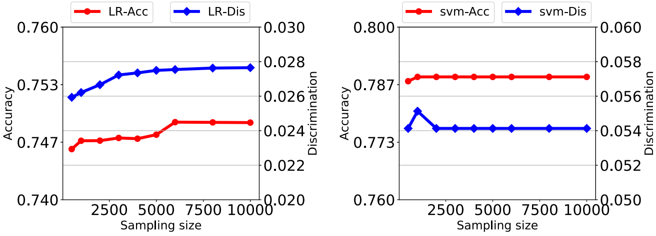
Fig. 4 shows the impact of sample size on accuracy and discrimination level with LR and SVM on three datasets. Overall, it is observed that accuracy increases quickly in the



(a) LR-Health (b) SVM-Health

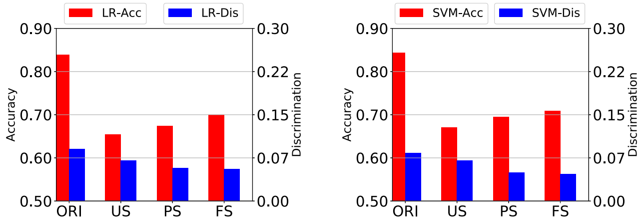


(c) LR-Bank (d) SVM-Bank

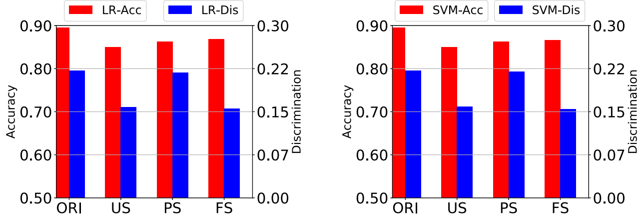


(e) LR-Adult (f) SVM-Adult

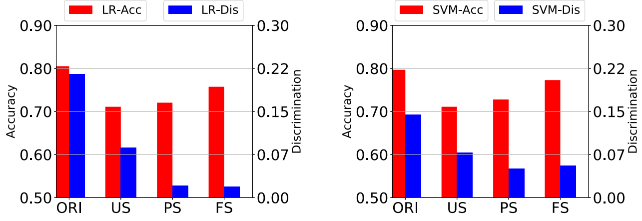
Fig. 4. The impact of sample size on accuracy (Red) and discrimination level (Blue) on (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. An increasing in the sampling size leads to an increase in accuracy and may help to reduce discrimination level.



(a) LR-Health (b) SVM-Health



(c) LR-Bank (d) SVM-Bank



(e) LR-Adult (f) SVM-Adult

Fig. 5. Comparison with original scheme (ORI), uniform sampling (US) and preferential sample (PS) with (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. With the fairness-enhanced sampling method (FS), discrimination decreases without much cost of accuracy or accuracy increases without much cost of discrimination.

early stages and then becomes stable as the sample size grows. This is because more data help to improve the generalization ability, but extra data do not help when the amount of data is enough to fit the model. Unlike accuracy, discrimination level depends on the amount of label noise that unlabeled data may bring when the sample size increases. For example, discrimination decreases in the Health dataset and increases a litter in the Bank dataset. This means that, with an increasing of sample size, little label noise is brought into the Health dataset, and consequently discrimination level decreases. Also, it is note that LR is more sensitive to sample size than SVM. The choice of sample size depends on the quality of the dataset and the training task requirement. Generally, a larger sample size can improve accuracy, reduce discrimination in bias and increase discrimination in noise.

5.1.5 *Comparison With Other Methods*

Fig. 5 shows the results from a comparison of our proposed FS method with and the other three schemes in terms of the accuracy and discrimination level on the three datasets. The training dataset of other methods is the original training dataset and the training dataset of our method is the new training dataset that consists of the original training dataset and pseudo labeled dataset (ρ = 1). The test dataset is the same. The results show that our method is able to push the discrimination to very low values while achieving a fairly high accuracy comparing with other schemes. Specifically, on the Adult dataset, the discrimination level under LR is around 0.215 with the original method and around 0.022 with the preferential sampling method, and the proposed FS method can decrease discrimination to 0.019 with a better accuracy than the preferred sampling method. This indicates that the proposed FS method is able to reduce the discrimination better than other methods.

5.2 Experiments on Synthetic Data

We first describe how to generate synthetic datasets and the goal of synthetic datasets is to show the effectiveness of our method in the discriminatory test dataset and fair test dataset. Here, the discriminatory test dataset refers to the test dataset whose data points are not equally presented in each group, and the fair test dataset refers to the test dataset whose data points are equally presented in each group. We show the distinct difference of discriminatory on two types of test datasets.

5.2.1 *Synthetic Data Setup*

We generate 22,000 binary class labels and a protected attribute a with a uniform random distribution, and assign a 2-dimensional feature vector to each label by drawing samples from two different Gaussian distributions: $p(x|y=1) = N([2;2],[5,1;1,5])$ and $p(x|y=-1) = N([-2;-2],[10,1;1,3])$. The size of each group in the synthetic dataset is roughly the same. Then we randomly sample 2,000 data points from

TABLE 1
Two Discriminatory Datasets Tested on the Discriminatory Test Dataset in ORI Method and the Proposed Fairness-Enhanced Method (FS) With LR and SVM

	Test with discriminatory test dataset						
	Method	Acc	Dis	G_{PP}	G_{UP}	G_{PN}	G_{UN}
LR	DA 1 (ORI)	0.8815	0.2705	183	586	626	605
	DA 2 (ORI)	0.8875	0.3642	104	628	642	626
	DA 1 (FS)	0.8825	0.2076	232	537	627	604
	DA 2 (FS)	0.8730	0.2890	159	573	642	626
SVM	DA 1 (ORI)	0.8825	0.2664	188	581	629	602
	DA2 (ORI)	0.8880	0.3724	102	630	649	619
	DA 1 (FS)	0.8825	0.2097	231	538	628	603
	DA 2 (FS)	0.8745	0.3130	149	583	655	476

We show accuracy (Acc), discrimination level (Dis) and the number of data points of each group in the discriminatory test dataset after classification.

TABLE 2
Two Discriminatory Datasets Tested on the Fair Test Dataset in ORI Method and the Proposed Fairness-Enhanced Method (FS) With LR and SVM

	Test with fair test dataset						
	Method	Acc	Dis	G_{PP}	G_{UP}	G_{PN}	G_{UN}
LR	DA1 (ORI)	0.8701	0.0484	438	556	492	514
	DA 2 (ORI)	0.8535	0.1018	376	618	483	523
	DA 1 (FS)	0.8790	0.0161	474	520	496	510
	DA 2 (FS)	0.8810	0.0062	471	523	483	523
SVM	DA1 (ORI)	0.8700	0.0483	441	553	495	511
	DA 2 (ORI)	0.8525	0.1118	372	622	489	517
	DA1 (FS)	0.8790	0.0168	474	520	496	510
	DA 2 (FS)	0.8775	0.0272	460	534	493	513

We show accuracy (Acc), discrimination level (Dis) and the number of data points of each group in the fair test dataset after classification.

the synthetic dataset as a fair test dataset, and split the remaining dataset randomly into two halves: one half is to be used as the labeled dataset and the other half with labels removed to serve as the unlabeled dataset.

Note that the synthetic dataset is a fair dataset, and the discriminatory dataset is generated by calibrating data points in the group G_{PP} based on the synthetic dataset. Discriminatory dataset 1 (DA 1) is generated by sampling 2,000 data points randomly in the group G_{PP} and data points do not change in other groups. Discriminatory dataset 2 (DA 2) is generated by sampling 3,000 data points randomly in the group G_{PP} and data points do not change in other groups. In each discriminatory dataset, we sample 2,000 data points as the discriminatory test dataset and the remaining as the training dataset.

5.2.2 Synthetic Data Tested With Discriminatory and Fair Datasets

Table 1 shows that our method is able to reduce discrimination level when training datasets have different discrimination levels. For example, more data points are classified into the Protected group with positive labels G_{PP} after implementing our method, and discrimination level of DA 1 reduces from 0.2705 to 0.2076 in LR. It is also note that accuracy does not decrease much with the proposed FS method. For example, accuracy of DA 2 reduces from 0.8825 to 0.8730 in LR.

We test the biased datasets with the proposed FS method on the fair test dataset with LR and SVM, and results are shown in Table 2. With the proposed FS method, discrimination level decreases and accuracy increases. More specifically, discrimination level decreases from 0.1018 to 0.0062 and accuracy increases from 0.8535 to 0.8810 in the DA 2. Discrimination level with the discriminatory test dataset is much higher than with the fair test dataset. We attribute this to the evaluation bias. Discriminatory dataset and discriminatory test data have the same data distribution, and thus the size of each group in the discriminatory test dataset is not equal. Even if the trained classifier is fair, the result may still be unfair. In real-world datasets, test datasets are sampled from the whole datasets and thus can contain evaluation bias.

5.3 Discussion and Summery

5.3.1 Discussion

We discuss on how the proposed FS framework is able to reduce discrimination in terms of discrimination decomposition into discrimination in bias, variance and noise. Discrimination in bias depends on the model choice. As we observe in the experiments, very broadly, LR can achieve a lower discrimination level than SVM. Discrimination in variance relates to the training data. Unlabeled data help to reduce discrimination in variance by increasing the size of training data. Ensemble learning helps to reduce discrimination in variance by averaging the training results from base models. An appropriate unlabeled data size, sample size and ensemble size in our framework is able to help reduce more discrimination in variance. Discrimination in noise depends on the quality of data. Training with unlabeled data may bring discrimination in noise. However, ensemble learning offsets this effect. When the same model is used, the benefit of unlabeled data in discrimination reduction depends on the impact of unlabeled data on discrimination in variance and discrimination in noise.

5.3.2 Summary

From these experiments, we see that the FS framework is able to reduce representation discrimination with a better trade-off between accuracy and discrimination. In the proposed FS framework, discrimination reduction in variance is usually more than the discrimination incurred by label noise. However, all the factors in the framework—model choice, unlabeled data size, ensemble size, sample size—each make their own particular contribution to increasing accuracy while ensuring fair representation.

6 RELATED WORK

In recent years, much research on fair machine learning has been undertaken. The following subsections summarize the three main streams of this work.

6.1 Pre-Processing Methods

Pre-processing methods eliminate the discrimination by adjusting the training data by ways of suppression, reweighing or sampling to obtain fair datasets before training [3], [11], [23]. Also, learning fair intermediate representations in the pre-process phase has received much attention. [12] was the first to open up fair machine learning by learning fair intermediate representations. The basic idea is that mapping the training data to a transformed space where as much useful information as possible is retained, but the dependencies between sensitive attributes and class labels are removed. Many researchers have subsequently studied fair representation learning with different methods, such as adversary learning [13], [14], [24], [25], [26]. These methods are based on using a classifier to predict sensitive attributes as adversarial components. The advantage of pre-precessing methods is that these methods can apply to all algorithms and tasks. Note that pre-processing approaches cannot be employed to eliminate discrimination arising from the algorithm itself.

6.2 In-Processing Methods

In-processing methods avoid discrimination with fair constraints [15] used regularizer term to penalize discrimination to enforce non-discrimination in the learning objective. [4], [27], [28] designed fairness constraints to achieve fair classification, where the fairness constraint is enforced by weakening the correlation between sensitive attribute and labels. In [29], [30], [31], the constrained optimization problem is formulated as a two-player game and fairness definitions are formalized as linear inequalities. Other recent work have a similar spirit to enforce fairness by adding constraints to the objective [32], [33]. The advantage of in-processing methods is that the level of fairness and accuracy can be controlled by the threshold of fairness constraints. However, fairness constraints are often irregular and need to be relaxed for optimization, and thus the solution may not be convergent. In addition, individual fairness can also be regard as in-processing methods [6], [34], [35].

6.3 Post-Processing Methods

A third approach to achieving fairness is post-processing, where a learned classifier is modified to adjust the decisions to be non-discriminatory for different groups. [5] proposed an approach to use of post-processing to ensure fairness criteria of equal opportunity and equal odds and subsequent work include [36], [37] However, it is not guaranteed to find the most accurate fair classifier [38], and requires test-time access to the protected attribute, which might not be available.

6.4 Comparison With Other Work

Existing fair methods focus on supervised and unsupervised learning, and these methods cannot be applied to SSL directly. As far as we know, only [39], [40] considered fair SSL. In [39], data is used to learn the output conditional probability, and unlabeled data is used for calibration in the post-processing phase. This method is to eliminate the aggregation discrimination, while the proposed FS method is to reduce representation discrimination. In [40], the proposed method is built on neural networks for SSL in the in-processing phase, and this method is to reduce measurement discrimination. In [11], representation discrimination is reduced by uniform sampling and preferential sampling, while in some cases not enough data in minority group can be sampled to generate a fair dataset. Our work make use of unlabeled data to form fairer datasets and theoretically analyze the discrimination via decomposition in bias, variance and noise. In our paper, we study the fair SSL based on label and unlabeled data in the pre-processing phase and our goal is to use labeled data to reduce representation discrimination, and in turn achieve a better trade-off between accuracy and discrimination.

7 CONCLUSION AND FUTURE WORK

In this paper, we use unlabeled data to achieve a better trade-off between accuracy and discrimination in the pre-processing phase. To achieve this, we developed a three-pronged strategy, where each component makes an important contribution to decreasing discrimination and/or improving the accuracy of the final predictions. Pseudo labeling in a semi-supervised setting exploits unlabeled data, on the premise that more training data is likely to reduce discrimination. A re-sampling method leads to multiple sampled fair datasets, and training on fairly-sampled will result in a fairly trained model. Lastly, ensemble learning is applied to improve the quality of the final predictions. A theoretical analysis and our experimental results show that our method delivers what it promises – unlabeled data is a viable option to achieve a better trade-off between accuracy and discrimination. Model choice, unlabeled data size, ensemble size and sampling size are factors that affect training results.

In future work, we intend to explore designs for fairness constraints that make use of unlabeled data to enforce fairness in the in-processing phase. Further, we have an assumption in this paper that labeled and unlabeled have the same distribution. However, this assumption may not hold in some real-world cases. Hence, another research direction is to how to achieve fair semi-supervised learning where labeled and unlabeled data have different data distributions.

ACKNOWLEDGMENTS

This work was supported by an ARC Discovery Project (DP190100981) from Australian Research Council, Australia; and in part by NSF under grants III-1526499, III-1763325, III-1909323, and CNS-1930941.

REFERENCES

[1] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[2] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. [Online]. Available: https://science.sciencemag.org/content/366/6464/447

[3] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2009, pp. 13–18.

[4] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, Apr. 20–22, 2017, vol. 54, pp. 962–970.

[5] M. Hardt *et al.*, "Equality of opportunity in supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3315–3323.

[6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innovations Theor. Comput. Sci. Conf.*, 2012, pp. 214–226.

[7] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," 2015, *arXiv:1511.00830*. [Online]. Available: https://arxiv.org/abs/1511.00830

[8] C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu, "Eliciting and enforcing subjective individual fairness," 2019, *arXiv: 1905.10660*. [Online]. Available: https://arxiv.org/abs/1905.10660

[9] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4066–4076.

[10] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 656–666.

[11] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2012.

[12] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 325–333.

[13] D. Madras, E. Creager, T. Pitassi, and R. S. Zemel, "Learning adversarially fair and transferable representations," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholmsmassan, Stockholm, Sweden, 2018, vol. 80, pp. 3381–3390.

[14] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, "Learning controllable fair representations," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, Apr. 16–18, 2019, vol. 89, pp. 2164–2173.

[15] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2012, pp. 35–50.

[16] M. Kleindessner, S. Samadi, P. Awasthi, and J. Morgenstern, "Guarantees for spectral clustering with fairness constraints," in *Proc. 36th Int. Conf. Mach. Learn.*, Jun. 09–15, 2019, vol. 97, pp. 3458–3467.

[17] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 3539–3550.

[18] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," 2019, *arXiv: 1901.10002*. [Online]. Available: https://arxiv.org/abs/1901.10002

[19] P. Domingos, "A unified bias-variance decomposition," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 231–238.

[20] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Representation Learn.*, 2013, vol. 3, Art. no. 2.

[21] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[22] X. J. Zhu, "Semi-supervised learning literature survey," Univ. Wisconsin-Madison, Madison, WI, *Tech. Rep. TR 1530*, 2005.

[23] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3992–4001.

[24] R. Feng, Y. Yang, Y. Lyu, C. Tan, Y. Sun, and C. Wang, "Learning fair representations via an adversarial framework," 2019, *arXiv: 1904.13341*. [Online]. Available: https://arxiv.org/abs/1904.13341

[25] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2018, pp. 335–340.

[26] D. McNamara, C. S. Ong, and R. C. Williamson, "Costs and benefits of fair representation learning," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2019, pp. 263–270.

[27] M. B. Zafar, I. Valera, M. Gomez Rodriguez , and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1171–1180.

[28] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2791–2801.

[29] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," 2018, *arXiv: 1803.02453*. [Online]. Available: https://arxiv.org/abs/1803.02453

[30] A. Cotter *et al.*, "Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals," *J. Mach. Learn. Res.*, vol. 20, no. 172, pp. 1–59, 2019.

[31] A. Cotter, H. Jiang, and K. Sridharan,, "Two-player games for efficient non-convex constrained optimization," 2018, *arXiv:1804.06500*. [Online]. Available: https://arxiv.org/abs/1804.06500

[32] J. Komiyama, A. Takeda, J. Honda, and H. Shimao, "Nonconvex optimization for regression with fairness constraints," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2737–2746.

[33] S. Aghaei, M. J. Azizi, and P. Vayanos, "Learning optimal and fair decision trees for non-discriminative decision-making," 2019, *arXiv: 1903.10598*. [Online]. Available: https://arxiv.org/abs/1903.10598

[34] T. Zhu and S. Y. Philip, "Applying differential privacy mechanism in artificial intelligence," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst.*, 2019, pp. 1601–1609.

[35] C. Dwork, C. Ilvento, and M. Jagadeesan, "Individual fairness in pipelines," 2020, *arXiv: 2004.05167*. [Online]. Available: https://arxiv.org/abs/2004.05167

[36] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2019, pp. 247–254.

[37] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias mitigation post-processing for individual and group fairness," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 2847–2851.

[38] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors," in *Proc. Conf. Learn. Theory*, 2017, pp. 1920–1953.

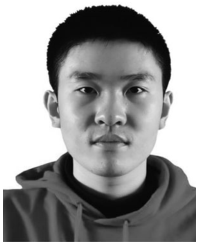
[39] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Leveraging labeled and unlabeled data for consistent fair binary classification," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 12 739–12 750.

[40] V. Noroozi, S. Bahaadini, S. Sheikhi, N. Mojab, and S. Y. Philip, "Leveraging semi-supervised learning for fairness using neural networks," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2019, pp. 50–55.

Tao Zhang received the BEng and MEng degrees from the Information Engineering School, Nanchang University, China, in 2015 and 2018, respectively. Currently, he is working toward the PhD degree in the School of Computer Science, University of Technology Sydney, Australia. His research interests include privacy preserving, AI fairness, and machine learning.

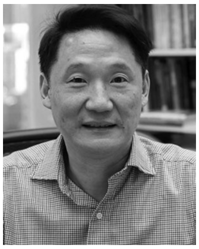
Tianqing Zhu received the BEng degree in chemistry and MEng degree in automation from Wuhan University, Wuhan, China, in 2000 and 2004, respectively, and the PhD degree in computer science from Deakin University, Geelong, Australia, in 2014. She is currently an associate professor with the Faculty of Engineering and Information Technology, School of Computer Science, University of Technology Sydney, Sydney, Australia. Before that, she was a lecturer with the School of Information Technology, Deakin University, Australia, from 2014 to 2018. Her research interests include privacy preserving, data mining, and cyber security.

Jing Li received the BEng and MEng degrees in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2015 and 2018, respectively. Currently, he is currently working toward the PhD degree in the Centre for Artificial Intelligence, University of Technology Sydney, Australia. His research interests include machine learning and privacy preserving.



Mengde Han received the master's degree from the Johns Hopkins University, Baltimore, Maryland. He is currently working toward the PhD degree at the University of Technology Sydney, Australia with a focus on Local Differential Privacy.

Wanlei Zhou (Senior Member, IEEE) received the BEng and MEng degrees from the Harbin Institute of Technology, Harbin, China, in 1982 and 1984, respectively, the PhD degree from The Australian National University, Canberra, Australia, in 1991, all in computer science and engineering, and the DSc degree (a higher doctorate degree) from Deakin University, Australia, in 2002. He is currently the head of School of Computer Science, University of Technology Sydney (UTS), Australia. Before joining UTS, he held the positions of alfred deakin professor, chair of Information Technology, and associate dean (International Research Engagement) of Faculty of Science, Engineering and Built Environment, Deakin University, Australia. His research interests include security and privacy, bioinformatics, and e-learning. He has published more than 400 papers in refereed international journals and refereed international conferences proceedings, including many articles in IEEE transactions and journals.



Philip S. Yu (Fellow, IEEE) received the BS degree in electrical engineering from National Taiwan University, Taipei, Taiwan, the MS and PhD degrees in electrical engineering from Stanford University, Stanford, California, and the MBA degree from New York University, New York. He was with IBM, Armonk, New York, where he was a manager of the Software Tools and Techniques Department, Thomas J. Watson Research Center, Ossining, New York. He is a distinguished professor of computer science with the University of Illinois at Chicago, Chicago, Illinois, where he also holds the wexler chair in information technology. He has published more than 1200 papers in peer-reviewed journals, such as the *IEEE Transactions on Knowledge and Data Engineering*, the *ACM Transactions on Knowledge Discovery from Data*, the *VLDBJ*, and the *ACM Transactions on Intelligent Systems and Technology* and conferences, such as KDD, ICDE, WWW, AAAI, SIGIR, ICML, and CIKM. He holds or has applied for more than 300 U.S. patents. His current research interests include data mining, data streams, databases, and privacy. He was a recipient of the ACM SIGKDD 2016 Innovation Award for his influential research and scientific contributions on mining, fusion, and anonymization of Big Data and the IEEE Computer Society 2013 Technical Achievement Award. He was the editor-in-chief of the *ACM Transactions on Knowledge Discovery from Data* (2011–2017) and the *IEEE Transactions on Knowledge and Data Engineering* (2001–2004). He is a fellow of ACM.