Analysis of Fair Learning without Sensitive Demographic Information

Yuji Roh [*1] **Hyungjun Yoon** [*1]

Abstract

Building a fair system is one of the critical issues in machine learning. Most fairness approaches assume the existence of sensitive demographic information (i.e., gender, race, or age), however, this assumption does not hold in many real-world applications. To address this issue, several recent studies try to build a fair system without sensitive demographic information. In this study, we first analyze one recent fairness algorithm that does not utilize group information and then suggest an improved version of the algorithm. The improved approach shows better fairness performances in various datasets.

1. Introduction

As machine learning systems begin to have a profound impact on society, the discussion deepened on whether these systems provide fair results for everyone. Recently, it has been found that machine learning models produce discriminatory results (even though they are not intended) for a specific group, and these unfair models become a more critical issue in high-stake applications such as finance and crime. To address this issue, researchers propose various fairness-aware algorithms (Zemel et al., 2013; Hardt et al., 2016; Zafar et al., 2017; Zhang et al., 2018; Roh et al., 2020; Jiang & Nachum, 2020) that mitigate any discrimination in the model.

One major assumption of most fairness approaches is the existence of sensitive demographic information (i.e., gender, race, or age), which may be not valid in many real-world applications. For example, in the United States, it is not allowed for financial institutions such as banks to request sensitive information about individuals (Chen et al., 2019). Also, even if we can access sensitive demographics, utilizing such information in the model training may be illegal in many applications (Barocas & Selbst, 2016). Several recent studies (Hashimoto et al., 2018; Mozannar et al., 2020; Lahoti et al., 2020) reveal the limitation of this assumption and suggest new fair algorithms that try to minimize the usage of sensitive demographic information. The proposed approaches show reasonable results in producing a fair output distribution over sensitive groups without using demographics. We provide more details on these approaches in Section 2.

Among these approaches, we focus on Lahoti et al. (2020), which is the most recent approach that does not use any knowledge of groups. The proposed algorithm successes to produce a fair model by adaptively giving more weights on less accurate areas in the model training. In the paper, such areas are called *computationally identifiable (CI) regions* (Hebert-Johnson et al., 2018), and the system finds and reweights the CI regions based on their adversarially reweighted learning (ARL) approach. In this study, we analyze and improve the ARL-based fair algorithm. The details of the original and the improved algorithms are described in Section 3.

The rest of the paper includes related work (Section 2), original and improved approaches (Section 3), and experimental results (Section 4) with analysis.

2. Related Work

In this section, we discuss noteworthy model fairness studies for a complete comparison.

To connect social goals and machine learning, many fairness definitions have been proposed (Verma & Rubin, 2018). Among the definitions, most model fairness studies focus on either individual fairness (Dwork et al., 2012) or group fairness (Barocas et al., 2019; Feldman et al., 2015; Hardt et al., 2016). Individual fairness aims to treat similar individuals similarly, while group fairness tries to ensure equal output distributions over different groups. Among the fairness metrics, this study focuses on group fairness by measuring the fairness level by comparing each group's area-under-the-curve (AUC) value.

As we discussed in the previous section, several recent papers propose new fair algorithms without using sensitive demographic information. Hashimoto et al. (2018) is the first work that focuses on when the sensitive group is un-

[*] Equal contribution on this project [1] School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea.

KAIST CS 774, AI & Ethics.

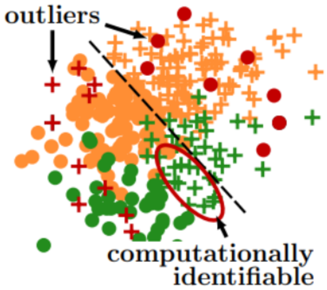


Figure 1. Computationally identifiable region (Lahoti et al., 2020). The shape(circle, plus sign) of each points means the label to be classified, and the color(orange, green) means the sensitive attribute they have. Assuming the dotted line is the classifier dividing labels, we can find a clustered region where a lot of plus signed points are misclassified. This region is defined as computationally identifiable region.

known. This study is based on distributionally robust optimization (DRO) (Sinha et al., 2017) to ensure fair results by equalizing risks over all distributions. Another line of study is Mozannar et al. (2020), which privatizes sensitive information by utilizing the locally ϵ-differentially private mechanism (Duchi et al., 2013). Since this approach still requires to access sensitive information to privatize the data, we do not have an immediate interest in this privacy-based approach. On the other hand, a more recent study (Lahoti et al., 2020) does not use any sensitive demographic information, while it improves the robustness on noisy labels rather than the DRO-based fair approach.

3. Approach

We now describe our main approaches. Our target algorithm called Adversarially Reweighted Learning (ARL) (Lahoti et al., 2020) is first discussed. We then provide the limitation of ARL and suggest the improved approach.

3.1. Original Algorithm

Our target algorithm is Adversarially Reweighted Learning (ARL) of (Lahoti et al., 2020). The key idea of ARL is to adaptively give more weights on less accurate areas, which is called *computationally identifiable (CI) region* (Hebert-Johnson et al., 2018), in the model training. The original paper assumes that the CI region can be determined with a binary function $f : X \times Y \rightarrow \{0, 1\}$ where the function f maps 1 if and only if $(x \in X, y \in Y)$ is in the CI region S. This concept is developed to capture high-loss regions from the learner as in Figure 1. In the figure where a classifier is dividing the data points which have label represented as different shapes, we note the region under the classifier where misclassified plus signs are clustered.

To implement their main idea the authors propose a minmax game between a *learner* and an *adversary* as in Figure 2. The learner is the primary classifier which optimizes θ for minimizing the original classification loss, while the adversary tries to maximize the shared loss with the primary by adding more weights to the computationally identified samples with a function mapping $f_\phi : X \times Y \rightarrow [0, 1]$. In other words, the learner and the adversary are competing with the minmax Equation 1 where $\lambda_\phi : f_\phi \rightarrow \mathbb{R}$ is the weight calculated from the adversary.

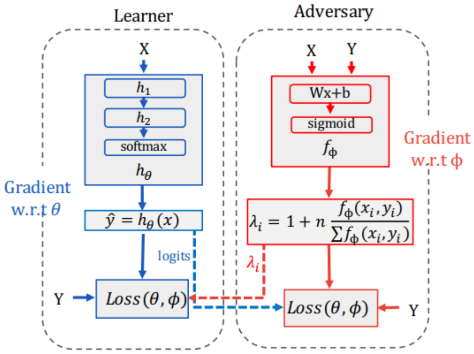


Figure 2. Overall structure of ARL (Lahoti et al., 2020).

$$J(\theta, \phi) = \min_{\theta} \max_{\phi} \sum_{i=1}^{n} \lambda_\phi(x_i, y_i) \cdot l_{ce}(h_\theta(x_i), y_i) \quad (1)$$

The weight parameter λ_ϕ is set as Equation 2, by considering following constraints: 1) restraining divergence to infinite, 2) positive value for stable behavior, 3) preventing fall to 0 to give meaningful contributions for each feature, and 4) inhibiting exploding gradients.

$$\lambda_i = 1 + n \frac{f_\phi(x_i, y_i)}{\sum f_\phi(x_i, y_i)} \quad (2)$$

To sum up, as in Figure 2, in ARL the learner and adversary alternately optimizes their own parameter by minimizing/maximizing the same weighted loss function Equation 1.

3.2. Limitation of the Original Algorithm

The main goal of the original algorithm is to make a fair classifier without demography with the assumption of the existence of CI regions. However, the competitive minmax game of ARL does not ensure the adversary to identify and weight the CI region as their objective. The authors did not provided any evaluation on the CI region identification in the paper. Moreover, the concept of using shared loss value (Equation 1) for the minmax game also brings the problem of unstable training. We suspect that these factors may cause the performance degradation in ARL and thus conduct an experiment to verify that the current architecture of ARL has limitation in capturing the CI region (Section 4.2).

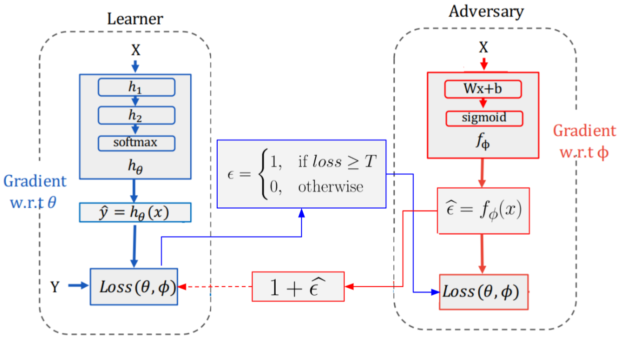


Figure 3. Improved ARL structure.

3.3. Improved Approach

To inhibit the problems of instability and uncertainty about the CI region identification, we propose a new form of ARL structure. We design improved ARL which gives weights in CI region without minmax game. The learner and the adversary do not share the same objective function, but the adversary has its own objective function for clearly identifying the CI region. In more detail, the adversary trains a binary classifier which infers high-loss samples from learner to 1, otherwise 0. We heuristically set the binary label ϵ for each sample by setting it as 1 if and only if the loss value from the learner output is larger than the pre-defined threshold T. The sigmoid output from the adversary $\widehat{\epsilon}$ is then used to give weights for the next round of the learner. To satisfy the weight constraints introduced in Section 3.1, we simply add 1 to $\widehat{\epsilon}$ for the weights. Figure 3 shows the overall structure of our proposal that simultaneously trains the learner and the adversary.

4. Experiments

We provide the experimental results showing the limitations pointed out in section 3.2, and the improvements on our revised ARL. First we demonstrate that our replicated system replicating original ARL works for the target datasets (AdultCensus (Kohavi, 1996), COMPAS (Angwin et al., 2016)) used in the original paper (Lahoti et al., 2020). Next, we show that the current ARL fails to capture the CI region, and our revised ARL overcomes the limitation in capturing the region. To effectively analyze the performance of CI region identification, we leveraged synthetic data (Zafar et al., 2017) which can be plotted in 2d space. Finally, we propose the improved performance on the target datasets by using our novel learning structure.

Data We use two real datasets and one synthetic data. Real datasets are utilized for reporting the replication results (Section 4.1) and performance of the improved approach (Section 4.3), and synthetic data is used for analyzing the limitations of the original algorithm and the validity of the improved approach.

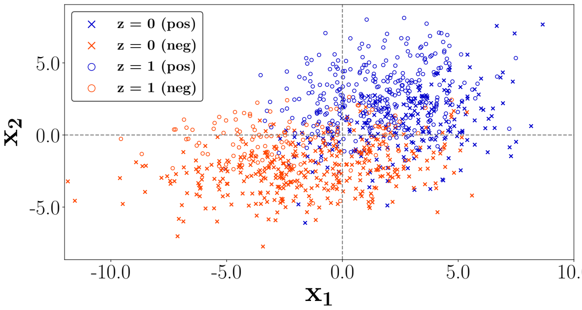


Figure 4. Synthetic data.

- AdultCensus (Kohavi, 1996): AdultCensus dataset has a variety of customer information and a binary label which represents whether each customer's annual income exceeds 50K dollars. In the experiments, we use GENDER and RACE as the sensitive attribute. We preprocess the dataset the same as in IBM 360 (Bellamy et al., 2018), and there are 43,131 examples in total.

- ProPublica COMPAS (Angwin et al., 2016): COMPAS dataset contains criminal information and a binary label which indicates whether each criminal re-offends. Similar to the AdultCensus dataset, we use GENDER and RACE as the sensitive attribute and make the same pre-processing as in IBM 360 (Bellamy et al., 2018). This dataset contains 5,278 examples.

- Synthetic data: As in Figure 4, we generate a synthetic data based on the proposals in (Zafar et al., 2017). It includes two non-sensitive attributes x_1 and x_2, a binary sensitive attribute z, and a binary label y. The (x_1, x_2, y) attributes are generated based on the two Gaussian distributions: $(x_1, x_2)|y = 0 \sim \mathcal{N}([-2; -2], [10, 1; 1, 3])$ and $(x_1, x_2)|y = 1 \sim \mathcal{N}([2; 2], [5, 1; 1, 5])$. The z attribute has the Bernoulli distribution $p(z = 1) = p((x_1', x_2')|y = 1)/[p((x_1', x_2')|y = 0) + p((x_1', x_2')|y = 1)]$ where $(x_1', x_2') = (x_1 \cos(\pi/4) - x_2 \sin(\pi/4), x_1 \sin(\pi/4) + x_2 \cos(\pi/4))$. We generate a total of 3,000 examples.

Evaluation Metrics The main goal of the original ARL in (Lahoti et al., 2020) is minimizing the gap across groups by observing the AUC for worst-case protected groups. The reason why they used AUC for the main metric is to consider robustness over class imbalance. Our replication purpose is to improve their performance with the same goal, so we follow the evaluation metrics proposed in their research. We report (1) weighted average, (2) macro average, (3) minimum value, and (4) value for the smallest protected group on the AUC over all protected groups.

Experimental Settings For the AdultCensus and COMPAS datasets, we conduct performance evaluation using the original ARL (Figure 2) and the improved ARL (Figure 3) designs respectively. For both architectures, the learner and the adversary follows the standard feed-forward network structure as in Lahoti et al. (2020). To simplify the task, we use a linear classifier for the learner. For adversary, we also adopt a linear structure based on the fact that the linear adversary worked the best in the original paper's evaluation. We note that in the experiment using synthetic data, since the artificial CI region we made formed non-linear region in the visualization space, we append one hidden layer with 32 units to give enough complexity for the adversary.

Our improved ARL requires one additional parameter T. T is defined as the threshold for deciding the labels which are feed into the adversary, representing whether the sample is correctly or not in primary. If we set lower T value then there would be more *misclassified-labeled* samples for adversary which results in wider range of CI region identification. Otherwise the captured CI region would be narrower. In our experiments, we set T as the value which outputs the best performance in each dataset.

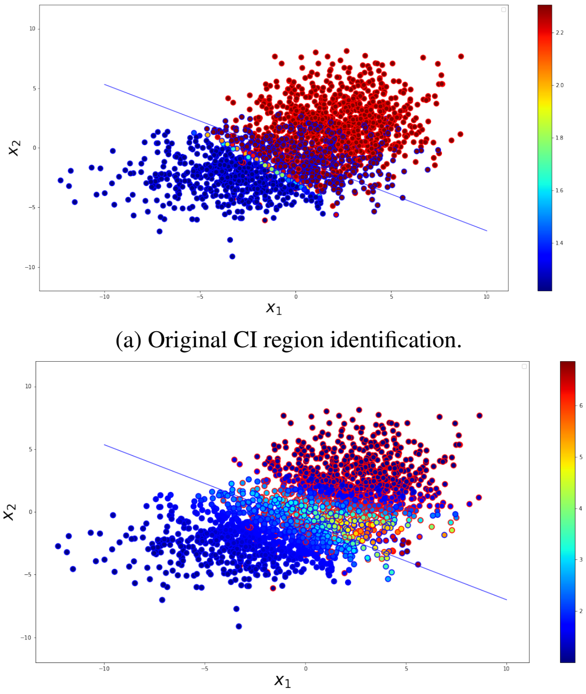
4.1. Algorithm Replication

We first reproduce the results of the original paper to check the validity of our replication. Logistic regression (LR) and ARL results in Table 1 show the performances of the non-fair classifier and the original ARL, respectively. The average AUC values (AUC avg and AUC macro-avg) indicate the overall performances of each algorithm. The fairness improvement can be observed via the changes in (1) AUC value of the group that has the minimum performance (AUC min) and (2) AUC value of the minority group (AUC minority). In the AdultCensus dataset, ARL achieves better AUC values than LR in both overall performance and fairness performance. On the other hand, in the COMPAS dataset, ARL rather worsens the performances of LR, especially in the minority group. This result implies that ARL may produce unpredicted performance degradation in several datasets. We remark that all observations for LR and ARL on both datasets are consistent with the original paper (Lahoti et al., 2020).

Although we succeed to replicate the original paper, the algorithm still has a performance issue. We suspect that the performance degradation on the COMPAS dataset is caused by the deficient capability in identifying the CI regions.

4.2. Analysis of the CI Region Identification

To analyze the CI region identification ability of the original ARL and the improved ARL, we visualized the output of adversary during the training phase. Under using a primary classifier which originally converges after epoch > 300 with learning rate 0.005, we changed the learning rate to 0.00005 and stopped training at epoch 300 before it converges ideal classifier. Since the classifier was not trained completely, there would be a bundle of samples that are misclassified. We defined the region these samples are gathered as CI region, and analyzed how the original ARL and the improved ARL captured it. Figures 5a and 5b show the results. The border color of each sample means the ground truth label and the line crossing the samples is the result of primary classifier. And the fill color of each point represents the output from the adversary. We could check that on the right upper region from the classifier line contains a lot of points with blue border so that the adversary should be able to capture the region. However, the original ARL's adversary failed to identify the region as in Figure 5a. It classified the whole region above. Meanwhile, as in Figure 5b, our improved ARL allocated high scores on the CI region we defined on the right upper side. We could check that the improved version outperforms in identifying CI region.



(a) Original CI region identification.

(b) Improved CI region identification.

Figure 5. Two versions of CI region identification. Each graph shows a temporal state of the learner and the adversary during training. Where the border color of each point is the ground truth label of them, the blue line crossing the points means the linear classifier of the learner. What we will focus on here is the fill color that represents the output of the adversary. The points which have colors close to red have higher confidence to be classified as computational identifiable. While Figure 5a shows the whole area above the learner as the red region, our improved ARL in Figure 5b shows clearly identified CI region in the right-upper side from the learner classifier line.

Table 1. Algorithm performances on the AdultCensus and COMPAS datasets.

Dataset	Method	AUC avg	AUC macro-avg	AUC min	AUC minority
AdultCensus	LR	0.698	0.695	0.688	0.688
	ARL	0.703	0.703	0.694	0.710
	Improved ARL	0.747	0.753	0.735	0.779
COMPAS	LR	0.677	0.639	0.602	0.623
	ARL	0.663	0.630	0.601	0.601
	Improved ARL	0.677	0.639	0.602	0.623

4.3. Improved Approach

We now discuss the performance of the improved approach. Table 1 shows that the improved ARL outperforms the original ARL in the AdultCensus dataset. For example, AUC min and AUC minority are increased to 0.735 and 0.779 from 0.694 and 0.710, respectively. In the COMPAS dataset, improved ARL at least does not degrade the LR's performance. Since the original ARL even worsens the LR, we believe that our new architecture improves the original version. We suspect that the core idea of ARL, which assumes that the CI regions exist, is invalid in the COMPAS dataset. Thus, the CI region-based designs (i.e., original ARL, improved ARL) seem to be difficult to work on this data.

5. Discussion

We improve the original ARL (Lahoti et al., 2020) by modifying the design for the adversary to effectively capture the CI regions. Although our improved ARL shows better experimental performances than the original ARL in the AdultCensus and COMPAS datasets, we also discuss two limitations of the current design to strengthen the system in the future.

One fundamental limitation of the ARL-based approaches is the practicality of the main assumption, i.e., the existence of the CI regions. As shown in Figure 1, the original paper assumes that the CI regions can be found by a binary function. However, there is no discussion on the validity of this assumption in real-world datasets. Since most real-world datasets have high feature dimensions, analyzing the validity of the assumption is not straightforward. Based on the fact that neither the original ARL nor the improved ARL can improve AUC performances, we guess that the COMPAS dataset may not have clear CI regions. The ARL-based approaches including our design may not work on the datasets without CI regions.

Also, we can improve how to select the hyperparameter T, which is one important aspect of our improved ARL. We currently choose the appropriate value of T via cross-validation. If the hyperparameter can be determined in a more systematic way, our design becomes more natural to apply in various applications.

6. Conclusion

Building a fair system is rising as an important issue in the ML community. There have been attempts to achieve fair learning with the sensitive group information, but in the wild it is hard to collect these sensitive attributes. Lahoti et al. (2020) is one of the most recent research addressing this issue. To produce a fair model without using the knowledge about groups, they build a learning system named ARL, which adaptively gives weights on the computationally identifiable regions. ARL adopts a min-max game between the learner and the adversary sharing the same objective as their key idea to find CI region.

We point out that their shared objective function does not ensure the adversary to capture CI region, and lacks robustness due to the competitive learning with the shared loss. To overcome the limitation, we propose improved ARL. In the improved ARL, an objective function only for detecting the CI region is newly assigned to the adversary, so that it can achieve weighing in the CI region without the previous min-max game. We show the improved CI region capturing performance by visualizing the identified regions using synthetic data. Furthermore, we demonstrate that our improved ARL achieves higher AUC compared to the original ARL in the AdultCensus and the COMPAS datasets. Based on the experimental results, we believe that our design outperforms the original design in identifying the CI regions and thus achieves higher performances in building a fair model without sensitive demographic information.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There's software used across the country to predict future criminals. And its biased against blacks., 2016.

Barocas, S. and Selbst. Big Data's Disparate Impact. *California Law Review*, 2016.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. Fairness under unawareness. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 2019. doi: 10.1145/3287560.3287594. URL http://dx.doi.org/10.1145/3287560.3287594.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *ITCS*, pp. 214–226, 2012. ISBN 978-1-4503-1115-1.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *KDD*, pp. 259–268, 2015. doi: 10.1145/2783258.2783311.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NeurIPS*, pp. 3315–3323, 2016.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *ICML*, pp. 1929–1938, 2018.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *ICML*, 2018.

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. In *AISTATS*, 2020.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, pp. 202–207, 1996.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. H. Fairness without demographics through adversarially reweighted learning, 2020.

Mozannar, H., Ohannessian, M. I., and Srebro, N. Fair learning with private demographic data, 2020.

Roh, Y., Lee, K., Whang, S., and Suh, C. FR-train: A mutual information-based approach to fair and robust training. In *ICML*, 2020.

Sinha, A., Namkoong, H., and Duchi, J. C. Certifying some distributional robustness with principled adversarial training. In *ICLR*, 2017.

Verma, S. and Rubin, J. Fairness definitions explained. In *FairWare@ICSE*, pp. 1–7, 2018. doi: 10.1145/3194770.3194776.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, pp. 962–970, 2017.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *ICML*, pp. 325–333, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *AIES*, pp. 335–340, 2018. doi: 10.1145/3278721.3278779.