Improving Machine Learning Fairness with Sampling and Adversarial Learning*

Jack J Amend and Scott Spurlock
Computer Science
Elon University
Elon, NC 27302
{jamend,sspurlock}@elon.edu

Abstract

Machine learning approaches learn models based on the statistical properties of training data. Learned models may be unfair due to bias inherent in the training data or because of spurious correlations based on sensitive attributes such as race or sex. This type of bias can lead to detrimental outcomes in important applications, including prison sentencing, credit scoring, and loan approvals. In this work, we perform a comparative study of techniques to increase the fairness of machine learning based classification with respect to a sensitive attribute. We assess the effectiveness of several data sampling strategies as well as of a variety of neural network architectures, including conventional and adversarial networks. Results are evaluated in terms of metrics measuring both classification accuracy and fairness. We find that model architecture and sampling strategy can both greatly affect metrics of fairness. We also find that there is no single best combination that should be used; the particular problem domain should drive the selection of neural network architecture and sampling strategy.

1 Introduction

Machine learning is becoming increasingly common in everyday life. Models are used to select ads to show users, recommend movies, and predict patient

outcomes. Advances in computing power, machine learning algorithms, and availability of training data have enabled the creation of models that can exhibit high accuracy, but are typically difficult to audit because of their complexity. Recently, many such models have been found to perpetuate societal biases. For example, a study by ProPublica found that a system that predicted recidivism scores was racially biased, predicting a higher likelihood of recidivism for Black individuals even when other factors were similar [1]. The frequency and disproportionate impact of this type of systemic bias motivate the necessity of finding ways to measure and mitigate bias in machine learning models.

In this paper, our focus is on *classification*, the most common application of machine learning, in which, given a data set of training examples and corresponding desired (ground truth) labels, a training process learns a model that can accurately predict the correct labels for given data examples. The training process seeks statistical patterns in the data that may be difficult or impossible for humans to identify. The resulting model is typically optimized to yield high levels of accuracy. We focus on artificial neural networks as the learning method most commonly employed in recent machine learning research and compare differing network architectures, particularly a vanilla "basic" network, and an adversarial network that optimizes competing objectives during training.

The training process is vulnerable to learning spurious correlations between attributes, particularly when the amount of data is limited. Sometimes these spurious correlations are harmless, e.g., learning that thin people always wear hats [9]. Learning is further vulnerable to codifying bias already present in the training data. These factors can result in models that are potentially detrimental, e.g., the association of Black individuals with higher rates of recidivism. Such factors as race and sex are of particular relevance to bias in models, and are often described as *protected* (or *sensitive*) attributes. Prior research has sought to quantify bias using several different criteria. Below we give definitions for two commonly used metrics. Following the notation of Zhang et al. [11], we use X, Y, and Z to indicate the input data, true label, and protected attribute, respectively. The model prediction for a given example is given by $\hat{Y} = f(X)$, where the function f is the learned model represented by a trained neural network. We indicate a particular value of the output variable, Y, by y, and of the protected variable, Z, by z.

- **Demographic Parity** measures that the predicted outcome is independent of the value of a protected attribute. $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z)$
- **Equality of Opportunity** measures that the predicted outcome is conditionally independent of the value of a protected attribute for *one par-*

ticular value of the outcome, which we indicate as 1.
$P(\hat{Y} = \hat{y}|Y = 1) = P(\hat{Y} = \hat{y}|Z = z, Y = 1)$

Our research evaluates strategies to reduce bias in learned models due to spurious correlations and biased training data that may have an adverse impact based on protected attributes. We conduct several experiments using the UCI Adult Data set [3] with the goal of predicting whether individuals belong to the high or low income group. We evaluate prediction accuracy with respect to the target variable and several fairness metrics with respect to the sex (male or female) attribute. Our experiments vary data sampling strategies as well as neural network architectures with the goal of addressing four research questions:

- **RQ1** Can a basic neural network achieve boosts in fairness metrics?
  In particular, we investigate whether basic network architectures can be effective when paired with an appropriate data sampling strategy.
- **RQ2** Which network architecture outputs the least biased predictions?
  We compare the results of training models with simple architectures as well as more complex, recently developed adversarial architectures.
- **RQ3** What is the best data sampling strategy to increase fairness?
  We compare several different strategies, including resampling to ensure the number of training examples is balanced over possible values of the sensitive attribute, of the desired label, and of both.
- **RQ4** Can we combine good architecture and data sampling to achieve better results?
  We evaluate the results of pairwise combinations of several sampling strategies and network architectures.

In the next section, we review recent work in the area of machine learning fairness. In Section 3, we describe our methodology, including data sampling strategies and neural network architecture choices, followed by a review of our experiments and results in Section 4. We conclude in Section 5 and offer some thoughts on potential directions for future work.

2 Related Work

Fairness in machine learning is becoming an active area of research. A recent survey focuses on the role that unbalanced training data can play in contributing to this issue, and groups work in the area into approaches that focus

on pre-processing the data and approaches that address the issue algorithmically [7].

Wang et al. evaluates several bias-reduction techniques in a computer vision context [10]. The authors propose training an ensemble of domain-independent classifiers (i.e., one classifier per possible value of a protected attribute). Interestingly, while this approach often outperforms a variety of alternatives, in some experiments, sampling with replacement to manually balance a training data set performs better. Oversampling techniques appear to be particularly effective when the data set is large (mitigating overfitting) and the amount of bias inherent in the data is smaller.

2.1 Adversarial Networks

A particular emphasis in much of the recent work has been on using specialized neural network architectures to help reduce bias in learned models. Adversarial learning seeks to optimize multiple neural networks with competing objectives. Typically, one network optimizes classification accuracy for a given model, while another network optimizes the ability to guess the value of a protected attribute given the classifier's output. By alternately training both networks, the goal is to learn a model that can predict with high accuracy while also exhibiting low levels of bias.

Several recent approaches [2, 4, 8, 11] propose to learn a mapping from input data to a new representation that removes bias from the source data. This learned representation is then suitable for learning unbiased classification models. The approach leverages an adversarial network seeking to predict a protected attribute based on the representation. Some work also finds that having balanced data sets in terms of the distribution of examples over the protected attribute is helpful in producing a fair model and that an adversarial approach allows for smaller numbers of training examples [2].

2.2 Fairness Metrics

There are many different metrics to measure fairness in a learned model, with new metrics being regularly proposed in the literature. Unfortunately, there is no consensus as to a single best approach to quantifying fairness, and there is generally a trade-off between model accuracy and various different fairness metrics. One recent study, which conducted a survey of the human perception of fairness of competing models in a hypothetical scenario, found that, while participants showed a slight preference for equalizing fairness over accuracy, they disagreed on how to measure it [6].

One measure to quantify fairness is equality of opportunity, introduced in recent work to remove bias from learned models [5]. Other common metrics

include demographic parity and equality of odds [11].

3 Methodology

In this section, we review our study's neural network architectures and data sampling strategies, as well as the data set used for evaluation.

3.1 Architecture

We experiment with four network architectures:

The **Basic** model is implemented as a simple fully connected 4-layer neural network that takes in the data, X, and outputs the predicted label $\hat{Y}$. This model serves as a baseline for comparison to the others.

The **Split** approach trains a separate Basic model for each of the possible protected attribute values, allowing each network to model a separate distribution. For a protected attribute like sex with two possible values, we train two independent basic models. At test time, each example is classified by the appropriate model.

The **CAN** (Classifier-Adversarial Network) architecture follows an adversarial learning approach, similar to several recent methods [2, 4, 11]. Adversarial learning works by pitting two competing neural networks against each other. The first, f, is the classifier, based on the Basic architecture described above, which attempts to predict the label, Y. The second network, g, uses the output from the classifier, $\hat{Y}$, to predict the protected attribute, Z. Training proceeds iteratively, alternately optimizing each network. After training, the networks reach an equilibrium, with the goal that the classifier performs with a high level of accuracy and the adversary performs poorly, near the level of random guessing in its ability to predict the protected attribute, thus limiting the correlation between the output of the classifier and the sensitive attribute.

The **CANE** model (CAN with Embedding), similar to CAN, trains competing classification and adversary networks. However, for CANE, the input to the adversary is augmented to include, in addition to $\hat{Y}$, the prediction from the classifier, the features from the penultimate layer of the classifier network. These features constitute an embedded, or lower-dimensional, representation of each input, X. They provide the adversary with more information, with the goal of helping to learn a less biased model. This variant of the CAN approach is actually more common in recent literature [2, 4, 11].

3.2 Data Sampling

Several recent approaches have focused on the impact of data sampling on fairness [5, 10]. To see how data affects the overall fairness of the model, we

Income	Female	Male
<=50K	9,592	15,128
>50K	1,179	6,662

Table 1: Counts of observations across income and sex in the highly unbalanced UCI Adult data set. Of 32,561 examples, there are many more low-income male observations (46.5%), while only 1,179 (3.6%) are high-income females.

compare 4 sampling approaches. No Sampling, **NS**, uses the data without modification, serving as a baseline. Sensitive Sampling, **SS**, resamples the training data so that the number of examples from each possible value of the sensitive attribute is the same (e.g., the same number of men and women). Label Sampling, **LS**, resamples the training data in a similar fashion with respect to the target variable, while Sensitive Label Sampling, **SLS**, equalizes the number of examples across each combination of sensitive attribute and target variable value.

3.3 Data

For our study, we selected the UCI Adult data set [3], which contains 14 continuous and categorical features including age, education, race, sex, and marital status, as well as an associated target variable indicating whether or not each individual's income is above or below $50K. This data set is well suited to our experiments because it is unbalanced in terms of the number of examples across both the sensitive attribute of sex as well as the target label. It has been shown to contain bias based on sex, and has been used in a variety of recent work on bias mitigation [2, 8]. As Table 1 shows, counts of observations across sex and income in the UCI Adult Data set are heavily skewed. Nearly two thirds of the observations are male and nearly three quarters of the observations are low-income. These disparities become even more apparent when looking at the counts for each combination of sex and income. Observations falling into both the high-income and female bins make up less than 4% of the entire data set.

4 Results and Discussion

In this section, we present our results and discuss findings for each of the four research questions.

Implementation The neural networks were implemented in Python using TensorFlow. For each architecture and sampling combination, models were trained for 100 epochs using the ADAM optimizer and a learning rate of 2e-4.

Model	Sampling	Acc. Overall	Acc. Female	Acc. Male	Parity Gap	Equality Gap -	Equality Gap +
Basic	NS	0.8479	0.9214	0.8116	0.1870	0.0808	0.1102
Basic	SS	0.8759	0.9399	<u>0.8119</u>	0.1732	0.0841	0.0446
Basic	LS	0.8744	0.9391	0.8097	0.1780	0.0880	0.0245
Basic	SLS	**0.8766**	0.9416	0.8116	0.1866	0.0936	<u>0.0069</u>
CAN	NS	0.8481	0.9192	**0.8129**	0.1613	0.0632	0.0257
CAN	SS	0.8742	0.9399	0.8085	0.1610	0.0776	0.0670
CAN	LS	<u>0.8764</u>	0.9413	0.8115	0.1631	0.0770	0.0513
CAN	SLS	0.8737	0.9402	0.8073	0.1458	0.0673	0.0899
CANE	NS	0.8444	0.9208	0.8066	<u>0.1375</u>	**0.0496**	**0.0030**
CANE	SS	0.8752	0.9394	0.8110	0.1647	0.0764	0.0218
CANE	LS	0.8732	0.9386	0.8078	0.1673	0.0820	0.0545
CANE	SLS	0.8582	0.9301	0.7862	**0.1301**	<u>0.0618</u>	0.0365
Split	NS	0.8428	0.9133	0.8080	0.1737	0.0712	0.0845
Split	SS	0.8532	**0.9478**	0.8064	0.1695	0.0896	0.0919
Split	LS	0.8539	<u>0.9468</u>	0.8080	0.1684	0.0876	0.0946
Split	SLS	0.8529	0.9447	0.8075	0.1748	0.0916	0.0816

Table 2: Experimental results with best value for each column bolded, second best underlined. For accuracies, higher is better; for gap metrics, lower is better. There tends to be a trade-off between better accuracy and gap metrics.

Metrics Table 2 lists the results for experiments with each of the models and sampling strategies. Results are averaged across multiple trials using 5-fold cross validation. For fair comparison, the same training-validation splits are used for each variant. Metrics include classification accuracy (overall and broken out for male and female) as well as fairness [2], based on the concepts of demographic parity and equality of opportunity defined in Section 1. The parity gap is calculated as the difference between probabilities of the model predicting high-income for the two sexes. The equality gap is the difference in probability of predicting each class, given the sex. This metric can be calculated for each of the target values, i.e., one for low and another for high income (Equality Gap - and +, respectively).

Question 1: Can a basic neural network achieve boosts in fairness metrics? Our results show that, for the basic network architecture, compared with not sampling, the other sampling strategies improve accuracy (SLS sampling yielded the highest overall accuracy of 87.66% across all experiments), but do not greatly improve fairness, with the exception of the high-income equality gap, which shows some of the lowest scores across all tests. Interestingly, the accuracy increase for SLS sampling comes primarily from more accurate classification of female examples, suggesting that this strategy improves the model's ability to generalize for this underrepresented set. This outcome is supported

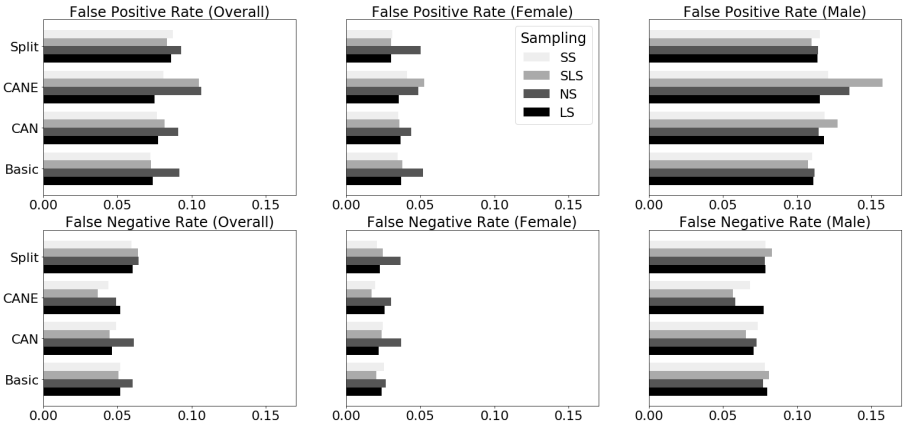

Figure 1: False positive (top) and false negative (bottom) rates for each sampling method grouped by architecture. The columns show (left to right) overall, female, and male rates, respectively.

by a closer look at classification error. Figure 1 breaks out errors in terms of false positive rate (FP), when the model incorrectly predicts high-income, and false negative rate (FN), when the model incorrectly predicts low-income. FP and FN are shown separately for overall, female, and male examples. For female examples, switching from no sampling to SLS causes the false positive rate to drop from 0.052 to 0.038, and the false negative rate from 0.027 to 0.021, while for male examples the error rate increases.

Question 2: Which network architecture outputs the least biased predictions? Compared to the basic model, all other model types result in some improvement in parity gap. This pattern was somewhat visible for the low-income equality gap, and less so for the positive equality gap. Overall, the adversarial architectures (CAN and CANE) produce models with better fairness metrics, and the CANE architecture unquestionably shows the best improvements to fairness metrics compared with the basic model. With no sampling, CANE results in the lowest equality gap (0.0496 and 0.0030) and second lowest parity gap (0.1375) across all experiments. The split model generally results in fewer improvements to fairness metrics, although combined with sampling strategies, does result in the highest accuracies for female examples in particular. Additionally, for false negative rate (Figure 1), we find a 17.47% decrease when using CANE with SLS vs. the CAN model with SLS. Compared to the basic model with SLS, false negative rate decreases by 27.43%.

Question 3: What is the best data sampling strategy to increase fairness? No single data sampling strategy improves all metrics across the board. For the adversarial models (CAN and CANE), no sampling (NS) generally results in the best parity and equality metrics, followed by SLS. For the basic and split models, parity gap is lowest for SS and LS sampling, with no clear-cut pattern for equality gap metrics. Figure 1 shows the impact of the type of sampling and model architecture on classification error rates. Most of the variability is due to female examples, with female false positive and negative rates exhibiting greater changes due to architecture and sampling choices. We theorize that this means that the models learn the distribution of females at varying levels based on the way the data is supplied and the model chosen. It is interesting to note that the false negative rates for CAN and CANE are similar for each of the resampling methods.

Question 4: Can we combine good architecture and data sampling to achieve better results? We find that overall, sampling strategies have more positive impact on the basic architecture. The adversarial architectures perform better from a fairness perspective with no sampling, although sampling can lead to accuracy improvements. In general, the results suggest a trade-off between classification accuracy and fairness, with improvements in one coming at the cost of reduction in the other. While there is no one clearly best combination of architecture and sampling, CANE with SLS provides the best scores on the fairness metrics. For a good compromise between accuracy and fairness, we note that CAN with LS scored second-highest in overall accuracy while achieving fairness scores near the median of all experiments.

5 Conclusion

In this paper, we evaluate the impact of data sampling and neural network architecture on classification accuracy and fairness metrics with a series of experiments on the UCI Adult data set. We find that sampling and architecture can both have important effects on classification results, but that no single combination of approaches yields top scores across all measures. Instead, there is a trade-off that allows an approach to be tuned to a particular domain where one metric may be more important than another. For example, a low false negative rate might be vital for medical diagnosis, while for credit scoring, a provably low bias might be required by law. For the future, further work investigating explicitly incorporating fairness metrics into neural network training may provide valuable improvements to learned models. We are also interested in learning generative models of data distributions to support data augmentation of under-represented examples.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016.

[2] Alex Beutel, Ed H. Chi, Jilin Chen, and Zhe Zhao. Data decisions and theoretical implications when adversarially learning fair representations. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[3] CL Blake and CJ Mertz. Uci repository of machine learning database, irvine, ca: University of california, 1998.

[4] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations*, 2016.

[5] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 2016.

[6] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Conference on Fairness, Accountability, and Transparency*, 2020.

[7] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019.

[8] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 2018.

[9] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840, 2013.

[10] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[11] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI Conference on AI, Ethics, and Society*, 2018.