DISCOVERY OF INTERSECTIONAL BIAS IN MACHINE LEARNING USING AUTOMATIC SUBGROUP GENERATION

Ángel Alexander Cabrera, Minsuk Kahng, Fred Hohman,
Jamie Morgenstern, Duen Horng Chau
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
{acabrera30, kahng, fredhohman, jamiemmt.cs, polo}@gatech.edu

ABSTRACT

As machine learning is applied to data about people, it is crucial to understand how learned models treat different demographic groups. Many factors, including what training data and class of models are used, can encode biased behavior into learned outcomes. These biases are often small when considering a single feature (e.g., sex or race) in isolation, but appear more blatantly at the intersection of multiple features. We present our ongoing work of designing automatic techniques and interactive tools to help users discover subgroups of data instances on which a model underperforms. Using a bottom-up clustering technique for subgroup generation, users can quickly find areas of a dataset in which their models are encoding bias. Our work presents some of the first user-focused, interactive methods for discovering bias in machine learning models.

1 INTRODUCTION

As machine learning (ML) becomes more advanced and established, it is being deployed to consequential domains ranging from facial recognition and recidivism prediction to automated credit scoring (Jordan & Mitchell, 2015). Along with this growth, serious fairness issues have been discovered where some ML models have better performance for certain demographic groups over others. These discoveries highlight the potential dangers of relying on learned, automated systems without fully understanding their realized impacts (Barocas & Selbst, 2016). When deploying ML models to the real world, it is important to know that they will treat different populations or subgroups similarly and will not produce significantly disparate results.

A major challenge for discovering fairness issues in ML models is that predictive inequity is often more pronounced in *intersectional* subgroups, and can be hidden when inspecting single features or aggregate accuracy metrics. In a striking example of this difficulty, Buolamwini & Gebru (2018) found that major facial recognition systems performed significantly worse at classifying the gender of darker skinned women than lighter skinned men, an imbalance that was not nearly as pronounced when looking at the features of skin tone or sex individually.

Discovering biases in ML models can be straightforward when looking at subgroups of a couple features, but it becomes much more challenging when attempting to discover bias in subgroups defined by potentially dozens of features (Kotkin, 2008). The number of generated subgroups grows exponentially as more features are considered, making it difficult to inspect every subgroup for fairness issues. Combined with the possibility that data scientists may not know what features they want to ensure fairness for, discovering intersectional bias becomes a daunting task.

Our novel techniques address the challenge of intersectional bias discovery by automating the subgroup generation process, directly providing users with potentially underperforming subgroups. Additionally, we enable users to discover and compare similar subgroups to investigate which features impact performance. Our contribution consists of two primary components:

1. **Subgroup Generation** Data instances are clustered into subgroups of various sizes. The entropy of each cluster's feature distributions is used to describe their feature makeup. Users can then explore the clusters to discover groups for which a ML model is encoding bias.

2. **Subgroup Comparison** The similarity between subgroups is calculated as the statistical distance between clusters. Once a user finds an interesting subgroup, they can compare it to similar groups to investigate feature and performance differences in a counterfactual manner.

There is ongoing research in both training models to mitigate intersectional bias and techniques for discovering potential biases. Kearns et al. (2019) introduce a training method that can address intersectional bias by enforcing fairness constraints on a set of known subgroups, and show the effectiveness of their method in practice. Our technique focuses on bias *discovery* and can be used to audit any model regardless of the architecture or training method used.

For intersectional bias discovery, Chung et al. (2018) propose a top-down method that can be used to find underperforming subgroups. They subdivide their dataset into more specific groups by considering more features until they find a subgroup with statistically significant loss. Lakkaraju et al. (2017b) take a different approach by using approximate rule-based explanations to describe subgroup outcomes. Our method is the first to take a bottom-up approach, using instances to create groups instead of feature values, allowing us to generate more diverse subgroups that can also generalize to any number of fairness metrics.

A survey from Hohman et al. (2018) investigated visual analytics tools and concluded that they can help users better understand and develop ML models, but there has been little work into systems specifically for ML bias and fairness. One of the first systems to address this gap is the *What If Tool* from Google (2018), that allows users to test different fairness constraints on a binary classification model. Our work presents one of the first user-focused techniques for exploring and discovering intersectional bias in general classification models.

2 SUBGROUP GENERATION

Discovering populations of a dataset that are underperforming for a given model is a challenging task, especially if the population is defined by various features. Often times users either do not know what features they want to ensure fairness for or there are simply too many subgroups to examine manually. Our method addresses these challenges by automatically generating subgroups of different feature combinations and sizes which have potential fairness issues.

To generate the subgroups, we begin by clustering the original data instances by their feature values. Since we do not have prior knowledge about cluster size or shape, we use the popular and efficient density-based clustering algorithm DBSCAN (Ester et al., 1996). The generated clusters represent groups of instances that have similar values and can be considered a subgroup or population. The notion of a "good" clustering is subjective, so we make the algorithm's hyperparameters available to the user so they can tune the clustering to balance the size, number, and uniformity of the groups. We are actively investigating how representative the generated clusters are of important subgroups, and which algorithms and hyperparameters produce the best groups.

Since the clusters are not explicitly defined by preselected features (e.g., the subgroup of Hispanic men), we need a way to describe each group of instances. To generate this description, we find which features are the most dominant in a cluster. We define a **dominant feature** as a feature of the cluster that is made up primarily of one value. For example, if for a given subgroup the feature *sex* is 99% female, we would call *sex* a dominant feature with value female. To measure the uniformity of features, we calculate the entropy of each feature distribution over its values. The closer a feature's entropy is to 0, the more uniform the feature is, making it more dominant in that subgroup.

Formally, suppose we have a set of features, $\mathcal{F} = \{f_1, f_2, ..., f_i, ...\}$, and each feature, f_i, has a set of possible values, $V_i = \{v_{i1}, v_{i2}, ...\}$. We calculate the *feature entropy* for the k-th subgroup and

i-th feature, $S_{k,i}$, as follows:

$$S_{k,i} = -\sum_{v \in V_i} \frac{N_{k,v}}{N_k} \log \frac{N_{k,v}}{N_k}, \tag{1}$$

where N_k is the number of instances in the k-th subgroup, and $N_{k,v}$ is the number of instances in the k-th subgroup with value v. For example, if all the instances of subgroup k have value $v_{3,1}$ (e.g., India), for the feature f_3 (e.g., native country), the feature entropy is 0 and f_3 is a *dominant feature* for the subgroup. We use the values of the top n dominant features to describe the group to the user. To explore the generated groups, the user can filter and sort them by size and different fairness metrics, allowing them to quickly find significant subgroups with anomalously low fairness metrics.

3 Subgroup Comparison

After having found a group with potential fairness issues, it can be useful to look at similar groups to investigate how their values and performance metrics differ. For example, two groups with one major feature difference and radically disparate accuracies may indicate that the feature is important for performance. This method is supported by recent research showing that counterfactual explanations have psychological backing and can be more useful than scientific explanations and approximations (Wachter et al., 2017). Our method provides potential counterfactual subgroups that can help explain the performance of a selected group.

To compute the similarity between subgroups we use the aggregate Jensen-Shannon (JS) divergence, a symmetric version of KL divergence, of all feature distributions between each pair of clusters. Summing the JS divergence of features gives us an aggregate measure of how similar different subgroups are. We calculate the total distance D between subgroups k and k' as follows, Where $G_{k,f}$ represents the value distribution of feature f in subgroup k.

$$D(k, k') = \sum_{f \in \mathcal{F}} \text{JS}(G_{k,f} || G_{k',f}) \tag{2}$$

Given two similar groups, their difference in values and fairness metrics can provide insight into the reasons for their underperformance. If two groups have large discrepancies in fairness metrics and only one or two of their features vary significantly, it may indicate that the anomalous features are important for performance. In contrast, if two groups have very similar performance, looking at their most identifying features can help users find more general subgroups with fairness issues.

4 Ongoing Work

4.1 Preliminary Experiments

Preliminary experiments using our techniques to analyze classification performance on the UCI Adult dataset (Dheeru & Karra Taniskidou, 2017) show promising results and compelling use cases. The UCI Adult dataset consists of 48,842 data instances, each with 14 features: 6 numeric and 8 categorical. The proposed classification goal is to predict whether a person makes more or less than $50,000 a year. To test our method, we trained a simple two-layer neural network using Adam as the optimizer (Kingma & Ba, 2014) and binary cross entropy loss that converges to an average training accuracy of around 84%. Lastly, we focus on the fairness definitions of negative and positive class balance and accuracy rates defined by Kleinberg et al. (2017). Our method allows users to use any fairness metric to analyze the generated subgroups.

Running our subgroup generation technique on the data creates 80 clusters from the instances, which we then sort by accuracy to find underperforming groups. While there are various small subgroups with accuracies ranging from 50% to 100%, subgroup 13 stands out as it has over 2,000 instances and a very low relative accuracy of 62.55% as seen in Figure 1. The dominant features of the subgroup describe it as made up of White men born in the USA who are married and have some college education. The user is then provided with various fairness metrics and information about the subgroup that can be used to investigate reasons for the low accuracy. A potential issue can be

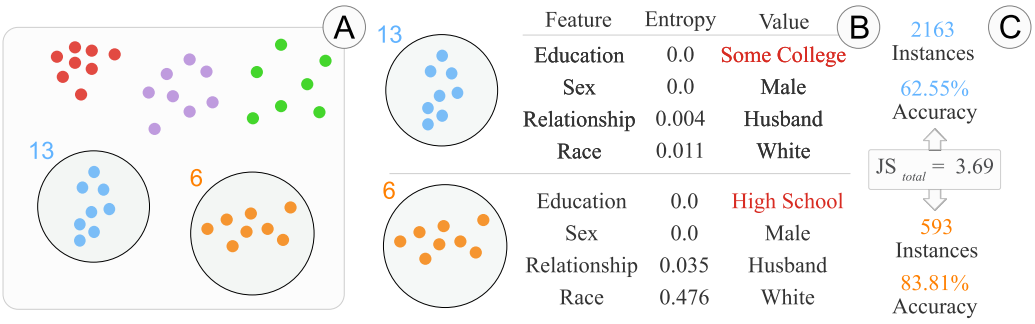

Figure 1: Using our technique on the UCI Adult Dataset we **(A)** Cluster instances into subgroups, then **(B)** Calculate subgroup feature entropy to find dominant features, and lastly **(C)** Investigate similar subgroups to discover value and performance differences.

quickly spotted in the class balance of the subgroup. Instances in the group are almost evenly split, with 1,216 people making under $50K and 947 making over $50K. The even split in labels can make it challenging for a classifier to discriminate between the two classes for instances in the subgroup.

To garner deeper insight about group 13's performance, we can look at its similar subgroups to find differences in values or performance. Group 6 is the third most similar group to 13, and is notable for its significant size of nearly 600 instances. Looking at their differences, we find that the two subgroups are nearly identical except for the education feature - people in subgroup 6 are high school educated whereas people in 13 are college educated. Investigating group 6 further, we see that its accuracy is significantly higher than group 13's, 83.81% compared to 62.55%. While this is a large gap in accuracy for two nearly identical groups, an explanation can be found when we see that in group 6, 497 people make less than $50K while only 96 make over $50k. This class imbalance allows the model to score a high accuracy of 83.81% for the subgroup by classifying every instance as making less than $50K. While this accuracy looks nominally better than group 13's, it is at the expense of having a 0% true positive rate.

Beyond finding underperforming subgroups, looking at similar groups allowed us to find that education appears to be correlated with the output class, impacting various fairness metrics like the accuracy and true positive rate. Our system allows users to quickly find insights about their data without necessitating any domain knowledge or manual generation of subgroups. This enables users to find issues much faster than could be done by manually generating and analyzing top-down subgroups.

4.2 Evaluation Plan

We plan on evaluating our tool through comparative, empirical user studies. It is currently possible to conduct intersectional bias discovery with existing tools, but it requires significant manual work and custom programming. In our user study, people with ML experience will be tasked with discovering biases in a given dataset and model output. Users will be randomly assigned a tool for producing their reports, either our system or an exiting ML exploration system as control groups.

In order to confirm that our approach can help users discover real biases, we plan on conducting our study using classic datasets and models with well-known fairness issues. These include the UCI Adult dataset presented above (Dheeru & Karra Taniskidou, 2017), the facial recognition dataset from the Gender Shades study (Buolamwini & Gebru, 2018), and more. To validate our system, we will measure which underperforming subgroups and important features the user is able to discover and how they rank the ease of use and discovery for our technique.

4.3 Future Directions

There are future promising directions in which we plan to extend our work. Interactive data visualization can significantly help users sort and explore a dataset that contains many groups found by our approach, as seen in existing visual analytics systems like the diagnostics system by Krause

et al. (2017) and the model investigation system ActiVis (Kahng et al., 2018). A next major step in ensuring machine learning fairness is suggesting and applying resolutions to biased models. There has been considerable research in this area, including the training methodology by Kearns et al. (2019) described above and the discovery of gaps in training data introduced by Lakkaraju et al. (2017a). Combining our approach with interactive visualizations that can also suggest potential resolutions could provide a thorough and powerful system for discovering and addressing bias in machine learning.

ACKNOWLEDGMENTS

This work was supported by a NASA Space Technology Research Fellowship, a Google PhD Fellowship, and NSF grants IIS-1563816 and CNS-1704701.

REFERENCES

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, New York, New York, 2018. ACM Press.

Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, and Steven Euijong Whang. Slice finder: Automated data slicing for model validation. *CoRR*, abs/1807.06068, 2018.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Martin Ester, Hans-Peter Kriegel, Jrg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pp. 226–231. AAAI Press, 1996.

Google. What if tool, 2018. URL `https://pair-code.github.io/what-if-tool/`.

Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Polo Chau. Activis: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics*, 24(1):88–97, 2018.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Conference on Fairness, Accountability and Transparency*, New York, New York, 2019. ACM Press.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017.

Minna J Kotkin. Diversity and discrimination: A look at complex bias. *Wm. & Mary L. Rev.*, 50: 1439, 2008.

Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. *Visual Analytics Science and Technology (VAST), IEEE Conference on*, Oct 2017.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017a.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017b.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2): 2018, 2017.