Contrastive Learning for Individual Fairness

Michael Xiang
michaelxiang@college.harvard.edu

Corwin Cheung
corwincheung@college.harvard.edu

***Abstract*—In recent years, contrastive learning has emerged as a powerful technique for training machine learning models, particularly due to its potential to learn rich representations of the data without explicit labeling. Contrastive learning operates by using positive and negative pairs to push similar data points closer together and dissimilar data points further apart within the context of the contrastive learner's embedding space. We realized that this principle is quite similar to the definition of individual fairness within the algorithmic fairness literature. Thus, this paper aims to explore how contrastive learning could potentially boost the individual fairness of a model. We found that in the COMPAS dataset our method outperforms the baseline in both fairness and accuracy.**

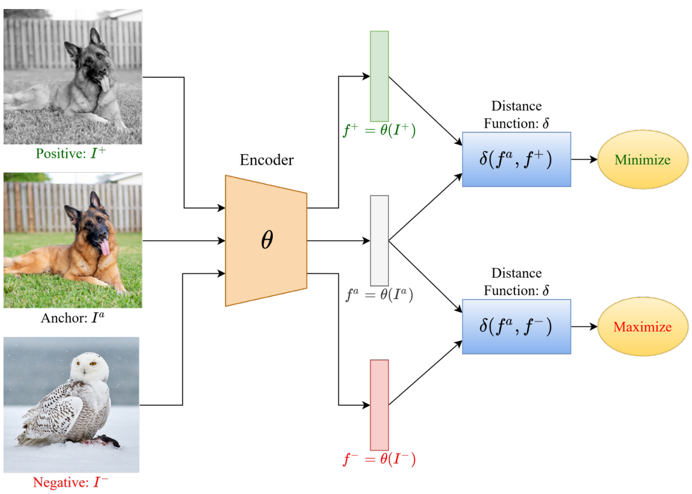

Fig. 1: This image shows how the contrastive learning process works. There are positive images that should be embedded close to the embedding of the anchor image and negative images that should be embedded farther apart. The encoder is trained with a loss function that aims to minimize the distance between the positive and anchor image embeddings while maximizing that for the negative and anchor image embeddings.

I. Introduction

Individual fairness, within the context of algorithmic fairness, is defined as treating similar individuals with respect to some task similarly in terms of their classification by the model [1]. One notable challenge, however, lies in finding an effective representation of the data. The way that the data gets introduced inherently introduces bias - certain features may "propogate" over into other features, e.g. someone may infer one's race from other factors, like one's neighborhood or one's salary from factors like zipcode. Thus, it is important that we do not mistakenly believe that removing sensitive features like race and sex from consideration of our data solves the individual fairness problem. We should actively ensure that people of different sensitive attributes are treated fairly.

To help with the problem of feature representation, we introduce the concept of contrastive learning. Within the context of training on images, contrastive learning operates by selecting some anchor image, performing an augmented version of that image to create a positive pair, and picking some other image within the dataset to act as a negative pair [2]. The idea is that, because the positive pair is formed from the anchor image, the embeddings for the two images should be very similar. And since the negative pair is formed from two different images, the embeddings should be vastly different in terms of distance in the embedding space. The images do not need to be labeled, and contrastive learning usually operates in an unsupervised learning environment. We observe that this training concept can be applied similarly to the goal of individual fairness - ideally, similar images should be mapped to the embedding space similarly, while different images should be mapped farther apart in the embedding space. Thus, it is our goal that we can exploit this training concept to optimize a model's individual fairness.

We note that, since contrastive learner uses an embedder, any model that uses the embeddings instead of the actual data loses some information because the embeddings are not lossless. Thus, there may be some loss in accuracy when we train models on embeddings generated by a contrastive learner rather than the original data. Within the algorithmic fairness literature, though, there are a lot of algorithms that have been developed recently that use the concept of boosting [3]. Boosting is a machine learning ensemble technique that combines multiple weak learners to create a strong learner. Weights are updated such that each weak learner fixes the mistakes of the the previous weak learner [4]. We hope that the loss in accuracy that comes with contrastive learning to be offset by the increase in model power with boosting when combining the two methods together.

A. Related Work

We build upon existing papers that explore contrastive learning and boosting within the realm of algorithmic fairness. Previous work in algorithmic fairness has yielded classic dataset example that we use to analyze our methods, specifically the COMPAS and the Adult datasets. We build upon existing approaches that attempt to both improve individual fairness via boosting and evaluate individual fairness through different metrics [3] [8] [9] [10]. Additionally, we further analyze contrastive learning to learn fair representations. Previously within the literature, such contrastive learners are only used on image or text, not tabular data [6], so we explore further how contrastive learning works on tabular data [5]. There has also been work done that uses labeled similar samples to learn individual fairness from the data. This approach is different than contrastive loss because contrastive loss has negative pairs and contrastive loss uses pairs we can generate synthetically from unlabeled data [7]. Further, those papers also seem to be interested in the fairness of contrastive learning in terms of group fairness metrics, not individual fairness [6], something that our project delves into.

II. Current Issues with Fair Models

We have already noted above that a huge problem within the algorithmic fairness community is the issue of fair representation of data. One of the fundamental goals of fair representation is to ensure that the features used by machine learning models to make predictions are not inadvertently encoding biases present in the data. Biases can manifest in various forms, including historical disparities, societal prejudices, and systemic inequalities, and they can permeate the data in subtle ways. These biases are reflected in the way that the outcome of data is measured, sample size bias for underrepresented groups, and decisions in processing and cleaning the data. Additionally, another issue with supervised learning in general is how expensive it can be to label datasets. In order to manually label examples that are similar to each other, quality controlled expert labelers are needed, which becomes quite time consuming for bigger datasets. This method is also not scalable to other datasets or tasks. Since contrastive learning is a form of unsupervised learning, it is less expensive and time intensive, and it is possible to use synthetic data (since we do not need labels) to train the model, which would both help dramatically cut costs and provides the model with much more data to work with. Thus, our goal for this project is to demonstrate that training classifiers on contrastively-learned embeddings rather than on the original data improves individual fairness, and that a boosting algorithm trained on the embeddings will retain the improvement in individual fairness while also improving accuracy.

III. Proposed Approach

As explained above, we hope to utilize contrastive learning to generate embeddings for the data, and then evaluate how those embeddings perform as training and testing data in comparison to the original dataset. We will compare this 3 ways. First, we will evaluate a logistic regression model on the original training data. Then, we will evaluate a logistic regression model trained on the embeddings of the data provided by the contrastive learner. Finally, we will evaluate a boosting model trained on the embeddings on the data provided by the contrastive learner. We will evaluate in terms of both accuracy and fairness.

Our datasets are tabular, and we will implement SCARF, a state of the art contrastive learner on tabular data [5]. SCARF works by taking the original training set and ”corrupting” it slightly to create synthetic positive pairs. For SCARF, we will adapt code from here. The idea is that, for a singular data point, it creates slight perturbations within that data point to generate a positive pair, and the synthetic perturbed data point should classify similarly to the original data point. We will tweak the training process, however, to train for individual fairness as well. For each training batch, we will create a copy of the batch and flip each of the protected attributes of each data point within the copied batch, so that the original and copied batch are similar except in terms of protected attributes. An example would be flipping race from a data point that is white to a copied data point with the same features except that the data point is non-white. We will train the SCARF model to treat the data points within the two batches as positive pairs. For individual fairness evaluation, we will follow the guideline established in the paper, individually fair gradient boosting, where we generate synthetic data and check to ensure that the ”flipped” version of the synthetic data (where everything is the same except for the protected attributes) is classified the same as the original synthetic data [3]. We define our fairness metric as the percentage of synthetic pairs where the flipped version was classified the same as the original. Our boosting model is the classical XGBoost model, which we import from a well-documented library here. Our pipeline for our code will be as follows. First, we will train SCARF, both with corruption and with positive pairs for fairness. Then, we will train a logistic regression model on the original data, a logistic regression model on the embeddings of the data produced by our modified version of SCARF, and a booster trained on the same embeddings. Finally, we will evaluate their accuracy using precision, F1, and recall scores as metrics, and we will evaluate their fairness by observing the percentage of times that the classifiers correctly predict the same output for synthetic positive pairs of data that only differ with respect to protected attributes.

IV. Experimental Setup

We will train our contrastive learner, SCARF, and then classification models on two datasets, the Adult dataset and the COMPAS dataset, both well-documented within the algorithmic fairness literature and found on AIF360 datasets [11]. We train for 2000 epochs total and batch size 128 for each dataset. We plotted the loss curves and obtained the classification reports for each dataset. The classification report contains our accuracy scores, which includes precision, recall, and F1 score. The fairness score, as mentioned before, is just the percentage of synthetic positive pairs predicted similarly by the model.

V. Results

We were able to demonstrate that the classifier trained on contrastively-learned embeddings either had similar or significantly better individual fairness scores compared to the classifier trained on original data, and that the booster trained on the embeddings had similar fairness scores and better accuracy metrics, compared to the original classifier trained on the embeddings. We have attached all charts in the appendix. For the adult dataset, we get the metrics for accuracy seen in figure 6. We observe that, overall, the logistic regression model trained on the original data has the best accuracy metrics while the booster trained on the embeddings has slightly better accuracy compared to the logistic regression model trained on the embeddings. Again, this is expected do to loss of information with embeddings. The accuracies are 0.84 for the logistic regression model trained with the original data compared .79 for both the logistic regression on embeddings and the booster on embeddings. We notice that the f1 score on the weighted average for the Booster on embeddings is slightly higher than the weighted average on the logistic regression on embedddings without the booster. In terms of fairness, all 3 models are on par with one another, with logistic regression on the original data slightly beating out the other two models by just over 0.01%, seen in table I. We also observe that the boosting model and logistic regression trained on embeddings have very similar fairness scores.

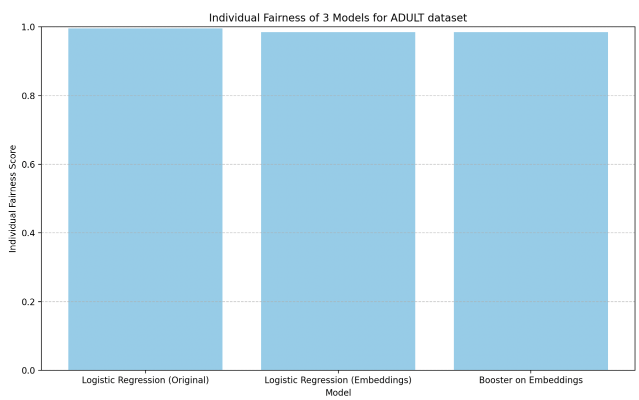

Fig. 2: Fairness Graph for the Adult Dataset

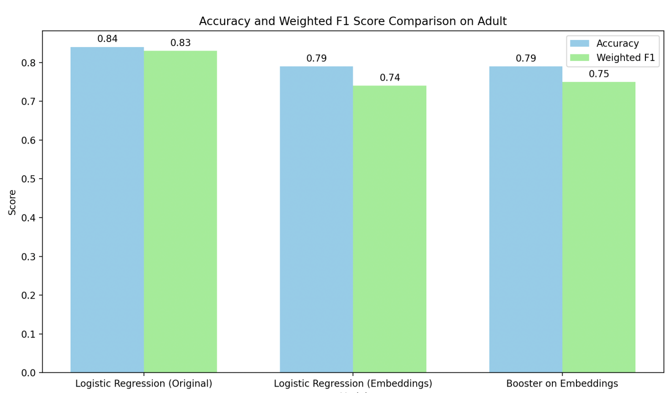

Fig. 3: Accuracy Graph for the Adult Dataset

For the COMPAS dataset, we get the metric for accuracy seen in figure 7. For accuracy, the booster on embeddings is the best of the three classifiers, slightly edging out the logisitc regression on original data. This is promising, especially because the booster has significantly higher fairness than the regression on original data. The logistic regression model trained on the embeddings still has the best fairness score, with the booster trained on the embeddings falling closely behind, less than 0.01 while the logistic regression model trained on the original data is 9% behind the logistic regression model trained on embeddings seen in table II. Thus with out method on the COMPAS dataset, we have that the embeddings significantly increase the individual fairness metrics and actually increase our accuracy as well.

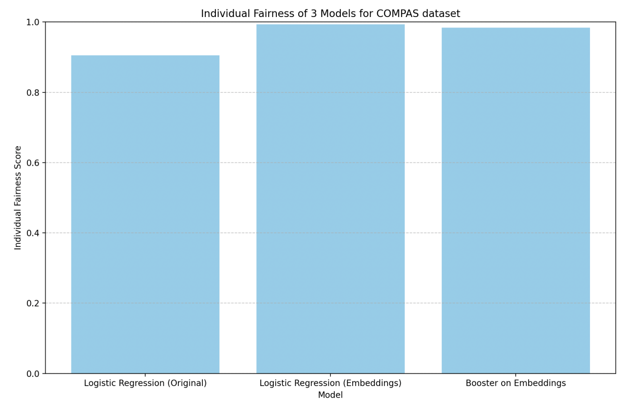

Fig. 4: Fairness Graph for the COMPAS Dataset

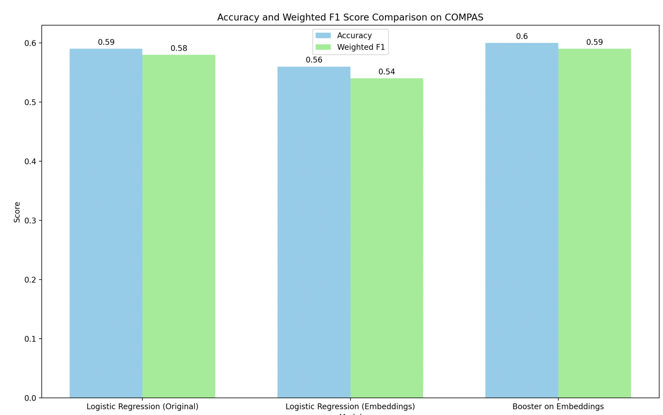

Fig. 5: Accuracy Graph for the COMPAS Dataset

Thus, to sum it up, we see that, for the adult dataset, the logistic regression model on the original dataset has the best accuracy and fairness scores, but booster model and the logistic regression model trained on the embeddings are very close behind in fairness. For the COMPAS dataset, the embeddings are superior to the original logistic regression in fairness and with the booster, we see slightly better accuracy than the original logistic regression. It is interesting to note that the Booster with embeddings sacrifices some individual fairness for a boost in accuracy. The logistic regression on embeddings without a booster has the best individual fairness score for COMPAS. Thus it seems like our embeddings were more effective with the COMPAS dataset than the Adult dataset. A hypothesis for this could be that the Adult dataset already exhibited very high individual fairness without the embeddings and thus there was little room to find improvements with the embeddings while the COMPAS dataset had a significantly worse individual fairness in the original logistic regressor and thus the embeddings had more room to be effective.

VI. Discussion of Results and Conclusion

We think our results showed promise of contrastive learning boosting individual fairness. Specifically, for the COMPAS dataset, we saw that the individual fairness was better for the logistic regression model trained on the embeddings rather than the one trained on the original data. For the adult dataset, we saw that individual fairness of the models trained on the original data was on par with that of the models trained on the embeddings. We think that this supports the idea that contrastive learning can improve individual fairness due to its training mechanism, in which we explicitly direct the embeddings to map similar people with respect to a task similarly in the embedding space regardless of their protected attribute. In terms of the tradeoffs with accuracy, we believe that this should be evaluated on a case by case basis. We also saw that the accuracy dropped when we trained on the embeddings rather than the original dataset. For us, this was to be expected, as embedding the data to a lower dimension loses information, so some loss of accuracy was to be expected. Boosting also slightly improved the accuracy of the model trained on embeddings while retaining the same levels of individual fairness, which we think supports the idea discussed that boosting can help increase accuracy without hurting fairness too much, due to its concept of combining weak learners and improving on the previous learner's failures. We believe that further work should be done to explore how these contrastive embeddings effect the accuracy in other types of datasets and whether the drop in accuracy depends on the size of the dataset. Through the COMPAS dataset, we conclude that it is possible for contrastive learning methods to improve individual fairness while retaining or even slightly improving accuracy due to its unique method of training on positive and negative pairs, and the booster trained on the contrastively-learned embeddings can be effective in making up for the possible information loss from the embeddings. We think further investigation into this phenomenon will significantly help validate our hypothesis, and that further training with, for instance, more epochs, may lead to more drastic results.

We have attached the code along with the paper for reproducibility, as well as the model weights at the repository here: https://github.com/mxiang04/Fair-Contrastive-Learning. Further code that details our modeling building process can be accessed here: https://github.com/CorwinCheung/Contrastive-Boosting We hope that this project nudges people to look into contrastive learning as a computationally inexpensive yet optimal way to train models to be more individually fair. Synthetic data provides an easy way to fit the model with a lot of data while preventing the need for labels, and contrastive learning provides a promising way to ensure that models align with the definition of individual fairness. With more time, we think it would be worth observing how our code works on other datasets and exploring different classification models to train our embeddings on. While the SCARF package is what we based our results and analysis on, we believe that further exploration into different schemes of creating positive and negative pairs for contrastive learning can yield promising results. With SCARF the negative pairs are the other examples in the batchs, however we hypothesize that there are other methods of picking negative pairs that should yield better results. By picking negative pairs of data points that are actually very different from each other, we could improve individual fairness metrics further since contrastive learning embeddings would separate more extreme data points. This grouping of negative pairs also offers a lens into analyzing contrastive learning's effect on group fairness through an embedding scheme similar to the one we are using for individual fairness.

VII. Group Contribution

Michael came up with the project idea to explore contrastive learning and boosting in the context of individual fairness. He also coded up the preprocessing required for the datasets, the evaluation code needed to assess accuracy, and the synthetic data generator to assess fairness. Corwin and Michael collaborated to code the contrastive learner which generates the embeddings and the booster which takes an ensemble of many learners. Corwin optimized the process of generating positive and negative pairs for the constrastive learner to take. Michael and Corwin worked together to write up the results, with both contributing to the interpretation of the results. Corwin coded up and created the visuals for the fairness scores and accuracy metrics. Corwin wrote up the related work section and most of the conclusion. Michael was responsible for most of the planning of the project.

References

[1] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12). Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 149, 1597–1607

[3] Alexander Vargo, Fan Zhang, Mikhail Yurochkin, and Yuekai Sun. 2021. Individually fair gradient boosting. *arXiv preprint arXiv:2103.16785.*

[4] Yoav Freund and Robert E. Schapire. 1999. A Short Introduction to Boosting.

[5] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. 2021. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv preprint arXiv:2106.15147*

[6] Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*

[7] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. 2020. Two Simple Ways to Learn Individual Fairness Metrics from Data. *arXiv preprint arxiv:2006.11439*

[8] Christina Ilvento. 2020. Metric Learning for Individual Fairness. *arXiv preprint arxiv:1906.00250*

[9] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, Guy Rothblum Proceedings of the 35th International Conference on Machine Learning, PMLR 80:1939-1948, 2018.

[10] Ivona Colakovic and Sašo Karakatič. 2023. Knowledge-Based Systems, Volume 280.

[11] Rachel Bellamy et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understadning, and Mitigating Unwanted Algortihmic Bias *arXiv preprint arxiv:1810.01943*

APPENDIX

A. Fairness Scores

TABLE I: Individual Fairness for Adult Dataset

Logistic Regression on Original Data	**Logistic Regression on Embeddings**	**Booster on Embeddings**
0.9960076162397887	0.9848289417111971	0.9842761501136293

TABLE II: Individual Fairness for COMPAS Dataset

Logistic Regression on Original Data	**Logistic Regression on Embeddings**	**Booster on Embeddings**
0.9046717171717171	0.9936868686868687	0.9842171717171717

B. Accuracy Metrics

	precision	recall	f1-score	support
0	0.87	0.93	0.90	21741
1	0.73	0.56	0.64	7200
accuracy			0.84	28941
macro avg	0.80	0.75	0.77	28941
weighted avg	0.83	0.84	0.83	28941

(a) Logistic Regression on Original Data

	precision	recall	f1-score	support
0	0.79	0.97	0.87	21741
1	0.73	0.24	0.36	7200
accuracy			0.79	28941
macro avg	0.76	0.60	0.61	28941
weighted avg	0.78	0.79	0.74	28941

(b) Logistic Regression on Embeddings

	precision	recall	f1-score	support
0	0.79	0.97	0.87	21741
1	0.72	0.24	0.36	7200
accuracy			0.79	28941
macro avg	0.76	0.61	0.62	28941
weighted avg	0.78	0.79	0.75	28941

(c) Booster on Embeddings

Fig. 6: Accuracy Results for the Adult Dataset

	precision	recall	f1-score	support
0	0.60	0.66	0.63	1957
1	0.57	0.51	0.54	1737
accuracy			0.59	3694
macro avg	0.58	0.58	0.58	3694
weighted avg	0.58	0.59	0.58	3694

(a) Logistic Regression on Original Data

	precision	recall	f1-score	support
0	0.56	0.74	0.64	1957
1	0.54	0.34	0.42	1737
accuracy			0.56	3694
macro avg	0.55	0.54	0.53	3694
weighted avg	0.55	0.56	0.54	3694

(b) Logistic Regression on Embeddings

	precision	recall	f1-score	support
0	0.60	0.72	0.66	1957
1	0.59	0.46	0.52	1737
accuracy			0.60	3694
macro avg	0.60	0.59	0.59	3694
weighted avg	0.60	0.60	0.59	3694

(c) Booster on Embeddings

Fig. 7: Accuracy Results for the COMPAS Dataset