An Empirical Study on Learning Fairness Metrics for COMPAS Data with Human Supervision

Hanchen Wang
Department of Engineering
University of Cambridge
hw501@cam.ac.uk

Nina Grgic-Hlaca
Max Planck Institute for
Software Systems
nghlaca@mpi-sws.org

Preethi Lahoti
Max Planck Institute
for Informatics
plahoti@mpi-inf.mpg.de

Krishna P. Gummadi
Max Planck Institute for
Software Systems
gummadi@mpi-sws.org

Adrian Weller
University of Cambridge
& The Alan Turing Institute
aw665@cam.ac.uk

Abstract

The notion of individual fairness requires that similar people receive similar treatment. However, this is hard to achieve in practice since it is difficult to specify the appropriate similarity metric. In this work, we attempt to learn such similarity metric from human annotated data. We gather a new dataset of human judgments on a criminal recidivism prediction (COMPAS) task. By assuming the human supervision obeys the principle of individual fairness, we leverage prior work on metric learning, evaluate the performance of several metric learning methods on our dataset, and show that the learned metrics outperform the Euclidean and Precision metric under various criteria. We do not provide a way to directly learn a similarity metric satisfying the individual fairness, but to provide an empirical study on how to derive the similarity metric from human supervisors, then future work can use this as a tool to understand human supervision.

1 Introduction

Bias in automated decision making systems has raised many concerns. One approach to address these concerns is to enforce individual fairness [Dwork et al., 2012], which requires treating similar people similarly. However, it is not straightforward to quantify the appropriate similarity of individuals. There have been some noteworthy subsequent works on this topic, such as [Zemel et al., 2013] and [Lahoti et al., 2019]. In our paper, we study the problem of learning an individual fairness metric from human annotated data. We leverage the intuition that human judgments might implicitly encode an underlying fairness metric they adhere to. We gather a dataset of human judgments about criminal recidivism risk predictions, and utilize this dataset to test the performance of several different metric learning algorithms on this data.

2 Related Work

Algorithmic Fairness. Most prior work on algorithmic fairness has focused on group notions of fairness, which require that protected groups as indicated by sensitive attributes receive similar treatment to others [Zafar et al., 2017b,a, Hardt et al., 2016]. On the other hand, individual fairness, introduced by Dwork et al. [2012], requires that similar individuals be treated similarly. To achieve this, one must first define a similarity metric that can be used to compare the individuals. This

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

similarity metric needs to be either given or learned from data. Some recent work on individual fairness [Speicher et al., 2018, Kearns et al., 2017, Liu et al., 2017] can be thought of as taking the former approach, since it implicitly incorporates the similarity metric in the optimization problem. Instead of specifying a similarity metric directly, we attempt to learn it from human annotated data.

Metric Learning. We leverage the rich literature on metric learning, which has been studied and applied in various domains, ranging from image processing [Fei-Fei and Perona, 2005] to recommendation systems [McFee et al., 2012]. In the past years, it has increasingly been applied on human annotated data, in order to model human notions of similarity [Tamuz et al., 2011]. In our work, we also take this approach. We refer the reader to Bellet et al. [2013] and Kulis et al. [2013] for a more in depth overview of the relevant literature on metric learning.

Learning Fairness Metrics. For literature on the metric learning from the fairness perspective, the recent work of Ilvento [2019] is closest to ours. They propose an approach for approximating an individual fairness metric from human judgments about the relative distance between inputs. While prior work did suggest that humans find it easier to make relative judgments than absolute ones [Stewart et al., 2005]. However, relative judgments are not as frequent in everyday decision making, for instance in recommendation systems, the collected data are about which articles the users have clicked, instead of their relative comparisons among these articles. In various scenarios, from granting bail to assigning social benefits, people are making absolute, and not relative judgments. Therefore, in real world applications, it may be more realistic to learn people's similarity metrics from their past (absolute) judgments, than to require them to make a set of new (relative) judgments. Nevertheless, since human might have noticeable uncertainty/noise on their predictions, thus we provide learning method and assessment benchmark based on the relative comparison from their point estimations instead of directly from predictions.

3 Methodology

3.1 Gathering Human Judgments

Scenario. In our experiments, we focus on the task of predicting criminal recidivism risk. We use a dataset related to the COMPAS tool – a tool used across the United States to help judges make bail decisions by predicting defendants' criminal recidivism risk on a 10 point scale [Angwin et al., 2016]. This dataset, gathered by ProPublica [Angwin et al., 2016], contains information about the recidivism risk predicted by the COMPAS tool, as well as the ground truth recidivism rates, for 7214 defendants who were arrested in Broward County, Florida, in 2013 and 2014.

Survey Instrument. To gather human judgments, we conducted an online survey in which we asked participants to estimate the likelihood of criminal recidivism of a fixed set of 200 defendants from the ProPublica dataset. To mitigate the effects of order bias [Redmiles et al., 2017], the defendants were shown in random order. For each defendant, participants were shown information about the defendant's demographics and criminal history, in the same format as in Dressel and Farid [2018] and Grgić-Hlača et al. [2019], and were asked to answer the following question: *How likely do you think it is that this person will commit another crime within 2 years?*.[1] Even though the COMPAS tool provides criminal recidivism risk predictions on a 10-point scale, our participants were asked to respond to this question using a 5-point Likert scale. We opted for this design choice in order to minimize the duration of our 200-question survey, since providing answers using 10-point Likert scales was found to be more time consuming than using 5-point scales [Matell and Jacoby, 1972].

Procedure. We recruited participants through the online crowdsourcing platform Prolific [Palan and Schitter, 2018]. Utilizing Prolific's advanced pre-screening options, we recruited 29 participants from the US who self-reported to have served on a jury. On average, the respondents took approximately 71 minutes to complete the survey, and were paid a base fee of 8.50 £. In order to increase response quality [Vaughan, 2017], we also provided a performance-based bonus. [2] As additional quality control measures, we discarded responses of participants who (i) did not respond to our 5 attention

[1] Participants were asked to answer three questions about each defendant. The two remaining questions were *Do you think this person should be granted bail?*, and *How confident are you in your answer about granting this person bail?*, but the results of this analysis go beyond of the scope of this paper.

[2] Performance-based payments have been found to increase the quality of responses in effort-responsive tasks [Vaughan, 2017]. Hence, in order to incentivize participants to provide high-quality survey responses, we

check questions correctly, or (ii) completed the survey in less than 45 minutes. After discarding the responses of these participants, our final sample consisted of 20 participants. These 20 participants had an average criminal recidivism prediction accuracy of 62.4%, close to the 62.1% and 60.2% that Dressel and Farid [2018] and Grgić-Hlača et al. [2019] reported that their participants achieved on the same task, hence providing additional evidence of the quality of the responses we gathered.

Dataset. The final dataset $\mathcal{D}$ consists of (i) the criminal recidivism risk scores $\mathcal{S}_j^i \in \{1, 2, .., 5\}$ provided by our 20 respondents i for 200 defendants j, as well as (ii) the COMPAS tool risk scores $\mathcal{S}_j^C \in \{1, 2, .., 10\}$ for 7214 defendants from the ProPublica dataset. The full dataset and a preview of the survey instrument, will be made publicly available once past anonymization requirements.

3.2 Distance Metric Learning

In our experiments, we evaluate the performance of several different Mahalanobis metric learning approaches on our criminal recidivism dataset $\mathcal{D}$. To cover a broad range of the learning methods, we considered one of each from the three learning paradigms discussed by [Bellet et al., 2013]: (i) fully supervised: Large Margin Nearest Neighbor(*LMNN*, [Weinberger et al., 2006]); (ii) weakly supervised: Mahalanobis Metric for Clustering(*MMC*, [Xing et al., 2003]); and (iii) semi-supervised: Least Squared-residual Metric Learning(*LSML*, [Liu et al., 2012]).

LMNN. [Weinberger et al., 2006] The fully supervised *LMNN* method can be directly applied on the labeled data provided by the COMPAS tool and our respondents. It attempts to minimize the distance between training instances and neighbors of same class, while keeping instances of other classes out of the neighborhood. During the implementation, we consider that instances with the same rating in our dataset belong to the same class. In other words, in this approach, even though our data consists of Likert scale ratings, we treat these ratings as categorical values, thereby losing some information.

MMC. [Xing et al., 2003] The weakly supervised *MMC* method is designed to work in scenarios when rich labeled data is not readily available, and takes pairwise relative comparisons as inputs instead. The algorithm maximizes the sum of pairwise distances between dissimilar pairs while keeping that of similar pairs relatively small. The learned metric by *MMC* can be constrained either in a diagonal form (weighted Euclidean) or a full one. We consider instances which have equal ratings in our dataset to be similar pairs, and the others to be dissimilar pairs. Again, as for *LMNN*, this approach treats our Likert scale data as categorical values, and leads to information loss.

LSML. [Liu et al., 2012] Unlike the *LMNN* and *MMC* methods, which can only utilize the 200 labeled instances, the semi-supervised metric learning algorithm *LSML* allows the use of the remaining ~7000 unlabeled instances from the ProPublica dataset as well[3]. It learns the metric from a set of triplet relative comparisons of the form "a and b are more similar than a and c". The triplet constraint set $\mathcal{C}$ is constructed as $\mathcal{C} = \{a, b, c | S_a \leq S_b + \sigma < S_c\}$, where $\sigma > 0$ and $S_{a,b,c}$ are the recidivism scores from our dataset. Unlike *LMNN* and *MMC*, *LSML* uses relative triplets, which allow us to capture more nuanced information available from our Likert scale judgments, such as "2 is closer to 3 than 5".

We implement an adapted version of the algorithm, by adding a trade-off coefficient α on the logdet regularization term. This adaptation allowed us to reduce the weight of the regularizer, thereby increasing the weight for satisfying the relative triplet constraints. As suggested by [Liu et al., 2012], we randomly subsampled $|N|$ training inputs, instead of utilizing the full set $\mathcal{C}$ whose size is $\mathcal{O}(|N|^3)$.

Procedure. Each metric (*LMNN*, *MMC*, *LSML*) is trained on 140 inputs and evaluated on 60 inputs. These 200 inputs were randomly selected. In our evaluation, we repeat this process 10 times and report the average results.

increased their bonus fee by $0.10 for each correct recidivism prediction, and decreased it by the same amount for each incorrect prediction.

[3]We ran the algorithm both with and without using the 7000 unlabeled COMPAS data points. The results were qualitatively similar and we report the results of running the algorithm without using the unlabeled data.

4 Experiments

4.1 Basic Analysis on the Dataset

As mentioned in 3.1, we have chosen a subset of 200 defendants with the similar demography distribution to the total 7214 defendants assessed by COMPAS tool. By analysing the collected recidivism predictions, bail decisions and decision confidence, we found that:

1) Bail decisions are divergent especially for defendants classified to be with a higher probability of recidivism. As we can see from table 1, the fraction of granting bails are highly different among the 20 human judges when the defendants are classified as "Most Likely" and "Likely" to recidivate.

Recidivism Jugement	Most Unlikely	Unlikely	Neither	Likely	Most Likely
max granted rate	100%	100%	100%	95.6%	76.5%
min granted rate	96.8%	77.9%	50.0%	0	0
mean granted rate	99.6%	96.6%	84.7%	43.1%	18.9%

Table 1: Divergence in the bail decisions under different recidivism likelihood judgments

2) Decision confidence among the 20 human judges are quite differently calibrated. We treat 'neither' as non-recidivate predictions, and compare the predictions with the ground truth, the general accuracy for all 20 human judges are not more than 70%, but for some human judges, the accuracy of their confident judgements are significantly higher, the results are shown in table 2.[4]

Accuracy	Overall	Judge 1	Judge 10	Judge 18
all predictions	0.624±0.027	0.635	0.580	0.580
confident predictions	0.649±0.061	0.828	0.750	0.714
unconfident predictions	0.619±0.037	0.594	0.526	0.544

Table 2: accuracy of predictions with different levels of confidence

4.2 Evaluation

In addition to the distance metric learning approaches discussed in section 3.2, we compare the three aforementioned metric learning approaches against two baselines: (i) the trivial baseline of the *Euclidean* metric (i.e., the ℓ_2 distance in the feature space), as well as (ii) the non-trivial, but naïve *precision* matrix (i.e., the inverse of the covariance matrix), which is a standard approach for removing correlation between features.

We utilize our dataset with 7:3 train test split, all the results are an average over 10 runs. We evaluate the performance of the learned metrics with respect to the three loss functions defined below:

Relative Comparisons. This loss calculates the percentage of relative comparison triplets (constructed as described in the previous section) from the test set $\mathcal{C}_t$ which violate the constraints $d_{\mathbf{M}}(a,b) \leq d_{\mathbf{M}}(a,c)$ for the distance metric $\mathbf{M}$, it is similar to the loss defined by [Hoffer and Ailon, 2015] but instead we assign equal weight to each instance:

$$L(\mathbf{M}) = \frac{H(d_{\mathbf{M}}(\mathbf{x}_a, \mathbf{x}_c) - d_{\mathbf{M}}(\mathbf{x}_a, \mathbf{x}_b))}{|\mathcal{C}_t|} \tag{1}$$

where $H(\cdot)$ is the Heaviside step function.

kNN L1. This loss calculates the ℓ_1 divergence between the test instance ground truth label and the weighted rating of its neighbors, defined by the metric $\mathbf{M}$:

$$L(\mathbf{M}) = \sum_{\mathbf{x}_i \in \mathcal{C}_t} |\hat{y}_i - y_i| = \sum_{\mathbf{x}_i \in \mathcal{C}_t} |\sum_j w_{ij} y_j - y_j| \tag{2}$$

[4] here the confident judgments are defined as the predictions with upper 50% prediction confidence

where the normalised weight factor w_{ij} is proportional to the inverse of the distance between i and its k nearest neighbors j: $w_{ij} \propto 1/d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)$.

kNN L2. This loss is similar to kNN ℓ_1 but instead calculates the ℓ_2 distance. Compared to the ℓ_1 norm, it introduces more penalty for the large predicted error.

4.3 Metric Learning on Human Judgments

For the collected human survey, we have implemented *LMNN, MMC, LSML* in Section 3.2 following the procedure described in section 3.3. The trade off coefficient α of the regularizer in our adapted *LSML* is set to 0.01, and the number of neighbors for calculating *kNN L1, kNN L2* is chosen to be five[5]. The results are shown in Figure 1.

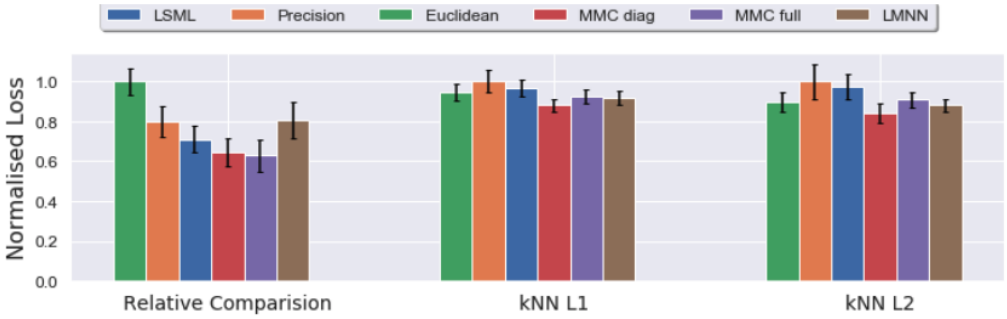

Figure 1: Performance of learned metrics based on the collected Prolific survey

From Figure 1, the kNN L1 and L2 loss of learned metrics are slightly better than the Euclidean and Precision probably due to there are still some sways especially when human make absolute ratings. For the triplet relative comparison loss, which does incorporate the relations of nearby ratings instead of treating them as categorical variables, the learned metrics has largely outperformed the Euclidean and Precision, which proves the effectiveness of our design.

4.4 Relative Comparisons on COMPAS

In this section we evaluate the sensitivity of our adapted *LSML* method to the hyperparameter σ, which controls the minimum required distance between inputs b and c. To this end, we compare the loss of the learned metric with the Euclidean metric on the much larger COMPAS dataset, instead of the human judgments we gathered. Recall Section 3.2, which describes how we construct the sets of triplet constraints based on the choice of σ. The loss based on the relative comparisons is then calculated on the resulting $\mathcal{C}_t$.

σ_t	**Euclidean**	**Ours**($\sigma = 0$)	**Ours**($\sigma = 2$)
0	0.40±0.037	0.39± 0.041	**0.38± 0.035**
2	0.40±0.027	0.39± 0.032	**0.37± 0.037**
4	0.35±0.031	0.31± 0.045	**0.30± 0.034**
6	0.31±0.033	**0.29± 0.060**	**0.29± 0.060**

Table 3: loss of learned metric by adapted *LSML* on the COMPAS system, here σ and $\sigma_t \geq 0$ represents the threshold values for constructing triplet constraints from the train, test set respectively.

Our learned metrics outperform the Euclidean distance by a large margin. As σ_t increases, the loss decreases for all three metrics, since the difference between b and c increases, hence providing us with less noisy data for learning the metric.

5 Discussion

In this work, we conducted a user study, in which we gathered a set of human judgments about recidivism risk, which we will make publicly available once past anonymization requirements. We

[5]We have tried varying these parameters, and they do not affect the high-level takeaways of our results

initiated work to examine various methods for learning an individual fairness metric from the human annotated data. Surprisingly, we observed similar performance across methods when considering predictive performance of ratings, though we saw differences when considering triplet consistency of unseen data. It will be interesting in future work to understand this better and consider the interplay between the metric-learning methods and the consistency of human ratings.

Acknowledgements

HW acknowledges support from Cambridge Trust CSC Scholarship, AW acknowledges support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via the CFI.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, 2016.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226, 2012. doi: 10.1145/2090236.2090255. URL `https://doi.org/10.1145/2090236.2090255`.

Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 524–531. IEEE, 2005.

Nina Grgić-Hlača, Krishna Gummadi, and Christoph Engel. Human decision making with machine assistance: An experiment on bailing and jailing. In *CSCW*, 2019.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6622`.

Christina Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1828–1836. JMLR. org, 2017.

Brian Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019.

Eric Yi Liu, Zhishan Guo, Xiang Zhang, Vladimir Jojic, and Wei Wang. Metric learning from relative comparisons by minimizing squared residual. In *2012 IEEE 12th International Conference on Data Mining*, pages 978–983. IEEE, 2012.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

Michael S Matell and Jacob Jacoby. Is there an optimal number of alternatives for likert-scale items? effects of testing time and scale properties. *Journal of Applied Psychology*, 56(6):506, 1972.

Brian McFee, Luke Barrington, and Gert Lanckriet. Learning content similarity for music recommendation. *IEEE transactions on audio, speech, and language processing*, 20(8):2207–2218, 2012.

Stefan Palan and Christian Schitter. Prolific.ac – a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 2018.

Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. A Summary of Survey Methodology Best Practices for Security and Privacy Researchers. Technical report, 2017.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248. ACM, 2018.

Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.

Jennifer Wortman Vaughan. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, 18:193–1, 2017.

Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2006.

Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.