Monitoring Fairness in HOLDA

Michele Fontana [a] Francesca Naretto [b] Anna Monreale [a] and Fosca Giannotti [b]
[a] *Università di Pisa, Pisa, Italy*
[b] *Scuola Normale Superiore, Pisa, Italy*

Machine learning (ML) models can solve complex predictive tasks, learning from a good training set. When data are human-generated, collecting them and moving them from one site to another might be difficult, due to privacy regulations, like the GDPR[1]. An interesting solution is Federated Learning (FL), proposed in [1] that trains a model across multiple devices, without moving their data and hence, keeping their data private. In the last few years, the FL approach has become popular: it has been adopted in query suggestion [2], next word prediction [3] and medical imaging [4]. There are two main approaches in FL: *cross-device*, in which the clients are mobile devices, and *cross-silo*, which deals with organizations. However, the privacy requirement is not the only ethical and legal principle to be guaranteed for the development of *trustworthy* AI systems. As required by the EU Commission with the AI Act regulation[2] other principles such as *fairness* and *transparency* are fundamental. In the literature, a lot of effort has been done in addressing privacy issues in FL systems [5,6,7] while the problem of potential unfair behavior of the ML models learned with a FL mechanism has been only marginally studied [8,9,10]. In this paper, we propose the study of a methodology for the monitoring of model fairness during the FL process.

Methodology One of the main differences between FL and the centralized setting relies on the fact that each node can have a different bias. When assessing the fairness in a cooperative setting it is thus important to understand *whether* and *how* each node could influence the others from the fairness point of view. For example, we might be interested in checking if a participant, starting from an unbiased set of data, could produce at the end an unfair local model. In other words, we need to assess the fairness according to several different settings, by varying the "degree of bias" in the federated data. In this work, we propose a methodology that allows to *monitor* and *assess* the fairness of the local models produced by HOLDA, a FL framework proposed in [11] during the execution of the training process. In particular, our monitoring procedure enables the study in FL of the tendency of biased participants to influence unbiased ones. The monitoring procedure acts in the following way. Whenever a client updates its internal state by replacing the previous best model with the new one, which exhibited a better generalization performance on the local validation data, the system assesses the fairness of this new model according to three different metrics: *Demography Parity* [12], *Equalized Odds* and *Equal Opportunity* [13]. These metrics are the ones that have been mostly adopted in the literature to assess the risk of unfair behavior.

[1] EU GDPR can be found at the following link: http://bit.ly/1TlgbjI.
[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

When dealing with FL methodologies, the first step that is required is to create the local datasets held by the local clients, starting from a unique centralized dataset. The main issue is that the datasets that are commonly used to benchmark problems about the risk assessment of unfair behavior, are typically small, such as the Adult dataset or the COMPAS ($48,000$ and $7,000$ respectively). Thus, they are not suitable for training good predictive models. In this work, we propose a strategy that allows adopting also small size datasets in FL tasks. Given a dataset D and a set of clients $c \in \mathscr{C}$, we first split D into $|\mathscr{C}|$ chunks, using random sampling with replacement. We precise that $|\cdot|$ denotes the cardinality function. We thus assign one chunk k to each client c and we use k as the training set for a generative model G. Once trained G, we query the generative model to sample new artificial data, which are thus drawn from a distribution that well approximates the one which underlies the records of D. As a consequence, these synthetic data can be used as the training set of the node c.

Preliminary Results We conducted preliminary experiments by training a neural network, having 200 neurons in the hidden layer, on the Adult dataset. We picked the "gender" attribute as the sensitive feature. We considered a federation made up of 10 clients, directly connected to the main server. In particular, half of the clients are unbiased, while the remaining ones have an extremely high bias w.r.t. the sensitive attribute. Our preliminary results suggest that biased participants may influence the unbiased nodes from the fairness point of view. In particular, the demographic parity is able to capture the difference in terms of bias between the two main groups of participants: the nodes without bias reach always a smaller value of this metric, meaning that their local models are fairer than the ones trained in the biased group of nodes. However, the value of the demographic parity for the unbiased participants is not around 0. Thus, their models tend to be slightly unfair despite the absence of any local bias. This idea is reflected much more by the equalized odds and the equal opportunity metrics: it might happen that the final model of an unbiased node could be more unfair than the local model of a biased one.

Future Works The ability of `HOLDA` to maintain in each iteration the best model enables to design some variants. Instead of only focusing on predictive performance for the local model selection, each node could also optimize the criterion of fairness of the model and maintain the fairest one. However, the choice of the best model can not be driven by just the fairness evaluation. In fact, it might happen that the fairest model could have bad generalization performance. As highlighted by the experiments, a node that is evaluated as perfectly fair could easily have an F_1 score which is lower than the performance of the best selected model. Thus, to come up with a model which is both well generalizing and fair, we need to solve a performance-fairness trade-off. One possible simple solution might be the following one: given two models M_1 and M_2, having similar performance, the algorithm has to select the fairest model among the two. Our future goal is to work on more and more sophisticated solutions that are able to take into consideration not only fairness but also other ethical values as privacy and transparency.

Acknowledgments. This work is supported by the EU H2020 project HumanE-AI-Net (Grant Id 952026) and CHIST-ERA grant CHIST-ERA-19-XAI-010, by MUR (grant No. not yet available), FWF (grant No. I 5205), EPSRC (grant No. EP/V055712/1), NCN (grant No. 2020/02/Y/ST6/00064), ETAg (grant No. SLTAT21096), BNSF (grant No. -06-2/5).

References

[1] Konečný J, McMahan HB, Ramage D, Richtárik P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. CoRR. 2016.

[2] Yang T, Andrew G, Eichner H, Sun H, Li W, Kong N, et al. Applied Federated Learning: Improving Google Keyboard Query Suggestions. CoRR. 2018.

[3] Hard A, Rao K, Mathews R, Beaufays F, Augenstein S, Eichner H, et al. Federated Learning for Mobile Keyboard Prediction. CoRR. 2018.

[4] Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence. 2020;2(6):305-11.

[5] Bhagoji AN, Chakraborty S, Mittal P, Calo S. Analyzing Federated Learning through an Adversarial Lens. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. vol. 97 of Proceedings of Machine Learning Research. PMLR; 2019. p. 634-43.

[6] Nasr M, Shokri R, Houmansadr A. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. 2019 IEEE Symposium on Security and Privacy (SP). 2019.

[7] Douceur JR. The Sybil Attack. In: Druschel P, Kaashoek MF, Rowstron AIT, editors. Peer-to-Peer Systems, First International Workshop, IPTPS 2002, Cambridge, MA, USA, March 7-8, 2002, Revised Papers. vol. 2429 of Lecture Notes in Computer Science. Springer; 2002. p. 251-60.

[8] Ezzeldin YH, Yan S, He C, Ferrara E, Avestimehr S. FairFed: Enabling Group Fairness in Federated Learning. CoRR. 2021;abs/2110.00857.

[9] Du W, Xu D, Wu X, Tong H. Fairness-aware Agnostic Federated Learning. In: Demeniconi C, Davidson I, editors. Proceedings of the 2021 SIAM International Conference on Data Mining, SDM 2021, Virtual Event, April 29 - May 1, 2021. SIAM; 2021. p. 181-9.

[10] Chu L, Wang L, Dong Y, Pei J, Zhou Z, Zhang Y. FedFair: Training Fair Models In Cross-Silo Federated Learning. CoRR. 2021;abs/2109.05662.

[11] Fontana M, Naretto F, Monreale A. A new approach for cross-silo federated learning and its privacy risks. In: 18th International Conference on Privacy, Security and Trust, PST 2021, Auckland, New Zealand, December 13-15, 2021. IEEE; 2021. p. 1-10.

[12] Calders T, Kamiran F, Pechenizkiy M. Building Classifiers with Independency Constraints. In: Saygin Y, Yu JX, Kargupta H, Wang W, Ranka S, Yu PS, et al., editors. ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA, 6 December 2009. IEEE Computer Society; 2009. p. 13-8.

[13] Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. In: Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R, editors. Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain; 2016. p. 3315-23.