Towards Bias Mitigation in Federated Learning

Yasmine Djebrouni

Towards Bias Mitigation in Federated Learning

Yasmine Djebrouni
University of Grenoble Alps (UGA) – LIG – LIRIS, France
yasmine.djebrouni@univ-grenoble-alpes.fr)

Abstract

Federated Learning (FL) provides better user data privacy, while allowing users to collaboratively solve a machine learning problem. However, FL can exacerbate the problem of model bias and unfairness, thus, resulting in segregative or sexist models. The objective of our PhD work is threefold: (i) Characterize the actual impact of FL settings on bias; (ii) Propose novel FL selection and aggregation methods for bias mitigation; (iii) Take into account antagonistic aspects such privacy, bias and robustness in FL.

1 Introduction

Machine learning is applied in many areas to extract knowledge and guide the decision making process, such as in search engines [13], recommendation systems [2], and disease diagnosis [12]. With the rapid growth of data, ML algorithms evolved from centralized to distributed solutions. In order to address data privacy issues, Federated Learning (FL) has emerged to allow a set of participants to collectively resolve a machine learning problem. Here, several data owners collectively learn from each others' data, without sharing their actual data. Thus, FL has applications in many areas such as health care, digital banking systems, etc.

However, FL can exacerbate the problem of bias and unfairness [1, 15]. Bias is a phenomenon that occurs when ML models produce unfair decisions due to the use of incomplete, faulty or prejudicial datasets and models. Bias may have serious consequences such as sexist segregation,illegal actions, or reduced revenues [3, 4, 16]. FL may exacerbate the problem of bias [3, 15], because of the decentralized nature of FL, where data distribution and size are particularly heterogeneous. Furthermore, data privacy constraints in FL do not allow the use of classical ML bias mitigation techniques [5, 17]. More precisely, our work aims to answer the following questions: (i) How to characterize the actual impact of Federated Learning on bias, *i.e.* to which extent do FL data distributions, FL models, FL selection, aggregation and robustness algorithms impact bias? (ii) What novel FL selection and aggregation algorithms could be proposed for bias mitigation? (iii) How to take into account privacy, bias and robustness in FL through a multi-objective approach, these objectives being usually antagonistic?

EuroDW'22, April 2022, Rennes, France
2022.

2 System Model and Problem Formulation

Federated Learning (FL). A FL system consists of N clients, each one holding its own data, and a server that orchestrates the overall FL learning in several rounds. At each round, the server selects a subset of m clients that participate to that learning round. The server sends the current version of the global model to the selected clients (in the first round, the server sends a randomly initialized model or a pre-trained model). Then, each participating client C_i trains the received model on its local data, and sends its local model updates θ_i to the server. Finally, the server performs a weighted aggregation of the received clients' model updates through FedAvg or other aggregation methods [14], to produce a new version of the global model θ as follows

$$\theta = \sum_{i=1}^{i=m} w_i.\theta_i \tag{1}$$

For simplicity, Eq. (1) assumes that the sum of weights w_i equals 1, although this could be generalized by dividing Eq. (1) by the sum of weights.

FL Bias Problem Formulation. We consider a binary FL classification model that learns over data where $(X_1, .., X_d)$ denote the features, Y denotes the class label, and $\hat{Y}$ is the classifier prediction result for a given data record. We consider two groups of data, namely a *privileged* group which prediction results have a given positive property p^* (*e.g.* people who earn a high salary), and an *unprivileged* group (*e.g.* people with a low salary). Let S be a sensitive feature which, for simplicity, we assume to be binary, $S \in \{a, b\}$ (*e.g.* a feature of race with two values, that are white or non-white). In a *biased* model, the value of S decides the membership of a data to either the privileged group (*i.e.* $\hat{Y} = p^*$) or to the unprivileged group, namely $S \in \{a = priv, b = unpriv\}$. Such a model does not provide group fairness [10]. With the latter, elements of the privileged group and unprivileged group have equal probability of having prediction results with a positive property, as formulated below:

$$Pr(\hat{Y} = p^*|S = priv) = Pr(\hat{Y} = p^*|S = unpriv) \tag{2}$$

Furthermore, in case of FL systems, the cause of bias of the global model can come from all or a subset of clients involved in a FL round. Thus, it is important to precisely determine the origin of bias in a FL system, to adequately mitigate it without hurting model quality.

Yasmine Djebrouni
University of Grenoble Alps (UGA) – LIG – LIRIS, France
yasmine.djebrouni@univ-grenoble-alpes.fr)

Bias Metric. A classical metric to quantify bias and measure group (un)fairness is *disparate impact* β [9], that is defined as follows:

$$\beta(\theta) = \frac{Pr(\hat{Y} = p^* | S = unpriv)}{Pr(\hat{Y} = p^* | S = priv)} \quad (3)$$

Here, we consider a model θ, and a small test set T of representative data of privileged and unprivileged groups, that is used to perform predictions with θ, and compute the bias metric β.

3 Proposed Bias Mitigation in FL

We propose a novel FL aggregation method that mitigates bias in the FL model. Roughly speaking, our method first estimates the contribution of each FL participant to the global bias of the FL model, and based on that information, reduces the impact of the participants causing bias on the overall aggregated model. In contrast to existing works [6–8, 18, 19], this method does not require additional information about client data distribution, which may leak sensitive information. Instead, it automatically monitors possible bias originating from each client, by measuring the bias metric (*e.g.* disparate impact) of each client's model updates, with a small test set T of representative data of each privileged and unprivileged groups. Thus, if $\beta(\theta_i) < \epsilon$, θ_i being the model update of client C_i, θ_i is considered as one of the causes of bias. Finally, one of the following model bias mitigation policies is applied.

***P-𝒲*: Pessimistic Weighted Aggregation in FL.** With this pessimistic policy, clients' model updates causing bias are simply ignored when producing the global aggregated model. That is, in Eq. (1), $w_i = 0$ for every θ_i for which $\beta(\theta_i) < \epsilon_u$; where ϵ_u is a threshold usually set to 80%, to ignore updates that are unfair with regard to unprivileged group. We can also consider the case where $\beta(\theta_i) > \epsilon_p$; with ϵ_p a threshold set to 120%, to ignore possible updates that are unfair with regard to privileged group.

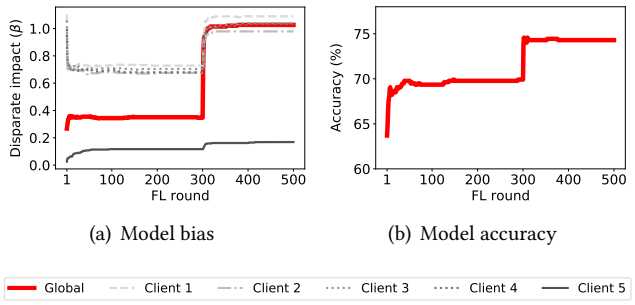
***IB-𝒲*: Inversely Bias Proportional Weighted Aggregation in FL.** In this bias mitigation policy, we consider new weights w'_i of model aggregation that are a function of usual FL weights w_i in Eq. (1) and the amount of bias induced by clients' model updates θ_i. More precisely, one of the following three situations are possible. If $\beta(\theta_i) = 1$, that means that θ_i fairly handles privileged and unprivileged groups. Otherwise, if $\beta(\theta_i) < 1$ (respectively $\beta(\theta_i) > 1$), that means that θ_i is unfair with regard to unprivileged group (respectively privileged group). Thus, we define the weights w'_i as follows:

$$w'_i = \begin{cases} w_i.\beta(\theta_i) & \text{if } \beta(\theta_i) \leq 1 \\ w_i/\beta(\theta_i) & \text{otherwise} \end{cases} \quad (4)$$

***MC-𝒲*: Mutually Cancelled Bias in FL Weighted Aggregation.** Instead of separately mitigating each client model updates θ_i like in *P-𝒲* and *IB-𝒲*, this new bias mitigation policy aims to take advantage of some biased clients' updates that can mutually cancel each other's bias. Thus, this results in an overall unbiased aggregated model, while enriching the model with additional data (*i.e.* for higher model quality and accuracy), instead of ignoring it like in *P-𝒲*, or reducing its impact like in *IB-𝒲*. Roughly speaking, *MC-𝒲* chooses the largest number k of selected clients m, that provides the highest aggregated contribution of updates and weights as defined in Eq. (4.

4 Preliminary Results

In the following experiment, we evaluate a FL system consisting of 5 clients, that train the logistic regression model on the Adult dataset [11], which contains information about employees, such as age, education level, gender, race, *etc.* The prediction task is to determine whether the income of a person is over \$50K or not. Here, the data sensitive feature is *sex*, and the target class is *salary* which is either high or low. Figure 1 presents the results of our experiment, where during the first 300 rounds, the classical FedAvg is used [14]. Then, following round 300, we apply our *P-𝒲* FL aggregation policy. We set ϵ_u to 60%. We observe that the global model is biased before round 300, with a disparate impact β of 0.53, and a lower model accuracy. After round 300, where client 5 is detected by *P-𝒲* as the origin of bias and, thus, ignored, bias is correctly mitigated, which also results in higher model accuracy.



(a) Model bias (b) Model accuracy

Figure 1. Before and after applying *P-𝒲*

5 Related Work

Several recent works study the problem of bias in FL. AgnosticFair [7] and FedFair [6], for instance, attempt to mitigate bias by introducing a fairness constraint into the global loss function. FairFL proposes a client selection method to improve fairness [19]. In FedFB, FL client pre-process their data to reweigh unprivileged data records [18]. In FairFed, FL clients locally measure the (un)fairness of their local data, share this information with the FL server, the latter takes into account this information to aggregate clients' updates [8]. Finally, state-of-the-art systems require additional client information about their data, *e.g.* the size of the unprivileged and privileged groups, which may leak private information.

6 Conclusion and Ongoing Work

We have described a novel FL aggregation method for model bias mitigation, with several policies, and presented preliminary evaluation results. Our ongoing work includes extensively characterizing the impact of FL system settings on actual model bias, implementing and evaluating the different bias mitigation policies through a multi-objective approach, evaluating our proposal in various real-world datasets and models.

References

[1] A. Abay, E. Chuba, Y. Zhou, N. Baracaldo, and H. Ludwig, *Addressing Unique Fairness Obstacles within Federated Learning*, (2021).

[2] S. B. Aher and L. Lobo, *Combination of Machine Learning Algorithms for Recommendation of Courses in E-Learning System Based on Historical Data*, Knowledge-Based Systems, 51 (2013), pp. 1–14.

[3] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al., *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*, arXiv preprint arXiv:1810.01943, (2018).

[4] A. Bogroff and D. Guegan, *Artificial Intelligence, Data, Ethics An Holistic Approach for Risks and Regulation*, University Ca'Foscari of Venice, Dept. of Economics Research Paper Series, (2019).

[5] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, *Optimized Pre-Processing for Discrimination Prevention*, in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., vol. 30, Curran Associates, Inc., 2017.

[6] L. Chu, L. Wang, Y. Dong, J. Pei, Z. Zhou, and Y. Zhang, *FedFair: Training Fair Models in Cross-Silo Federated Learning*, arXiv preprint arXiv:2109.05662, (2021).

[7] W. Du, D. Xu, X. Wu, and H. Tong, *Fairness-Aware Agnostic Federated Learning*, in Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), SIAM, 2021, pp. 181–189.

[8] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and S. Avestimehr, *FairFed: Enabling Group Fairness in Federated Learning*, arXiv preprint arXiv:2110.00857, (2021).

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, *Certifying and Removing Disparate Impact*, in proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

[10] M. Hardt, E. Price, and N. Srebro, *Equality of Opportunity in Supervised Learning*, Advances in neural information processing systems, 29 (2016), pp. 3315–3323.

[11] R. Kohavi et al., *Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid.*, in Kdd, vol. 96, 1996, pp. 202–207.

[12] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, *Machine Learning Applications in Cancer Prognosis and Prediction*, Computational and structural biotechnology journal, 13 (2015), pp. 8–17.

[13] A. McCallumzy, K. Nigamy, J. Renniey, and K. Seymorey, *Building Domain-Specific Search Engines With Machine Learning Techniques*, in Proceedings of the AAAI Spring Symposium on Intelligent Agents in Cyberspace. Citeseer, Citeseer, 1999, pp. 28–39.

[14] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, *Communication-Efficient Learning of Deep Networks From Decentralized Data*, in Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.

[15] H. B. McMahan et al., *Advances and Open Problems in Federated Learning*, Foundations and Trends® in Machine Learning, 14 (2021).

[16] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, *Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5310–5319.

[17] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, *Learning Fair Representations*, in Proceedings of the 30th International Conference on Machine Learning, S. Dasgupta and D. McAllester, eds., vol. 28 of Proceedings of Machine Learning Research, Atlanta, Georgia, USA, 17–19 Jun 2013, PMLR, pp. 325–333.

[18] Y. Zeng, H. Chen, and K. Lee, *Improving Fairness Via Federated Learning*, arXiv preprint arXiv:2110.15545, (2021).

[19] D. Y. Zhang, Z. Kou, and D. Wang, *FairFL: A Fair Federated Learning Approach to Reducing Demographic Bias in Privacy-Sensitive Classification Models*, in 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 1051–1060.