PRIVACY PROTECTION IN LEARNING FAIR REPRESENTATIONS

Yulu Jin and Lifeng Lai

Department of ECE, University of California, Davis
Email:{yuljin,lflai}@ucdavis.edu

ABSTRACT

In this paper, we develop a framework to achieve a desirable trade-off between fairness, inference accuracy and privacy protection in the inference as service scenario. Instead of sending raw data to the cloud, we conduct a random mapping of the data, which will increase privacy protection and mitigate bias but reduce inference accuracy. To properly address the trade-off, we formulate an optimization problem to find the optimal transformation map. As the problem is nonconvex in general, we develop an iterative algorithm to find the desired map. Numerical examples show that the proposed method has better performance than gradient ascent in the convergence speed, solution quality and algorithm stability.

***Index Terms*—** statistical inference, fair representation, privacy protection, iterative algorithm.

1. INTRODUCTION

As the number of IoT devices being introduced in the market has increased dramatically, inference as service (IAS) has been widely used in many sensitive environments to make decisions in the cloud [1]. In IAS, devices will send data to cloud and machine learning algorithms can be run on the cloud providers' infrastructure where training and deploying machine learning models are performed on cloud servers. However, two important issues, namely data privacy and fairness, need to be properly addressed.

Data privacy governs how data is collected, shared and used. In the IAS scenario, the devices send raw data to the cloud, which brings issues such as whether or how data is shared with third parties and how data is legally stored. In our recent work [2], we have addressed such privacy issue by transforming raw data by a carefully designed privacy-preserving mapping and sending the transformed data to the cloud. We show in [2] that such a transformation can provide a desirable trade-off between inference accuracy and privacy protection. There are also other interesting works on the privacy-utility trade-off while using mutual information as the utility measure [3, 4].

While these works address the privacy issue, they do not take the fairness issue into consideration. The main purpose of the fairness consideration in learning system is to ensure that the inference decisions do not reflect discriminatory behavior toward certain groups or populations. With the popularity of machine learning over the past decades, and their wide spread applications in many scenarios that have a direct effect in our lives, fairness constraints have become a huge issue for researchers [5]. There are at least two potential sources of unfairness in machine learning outcomes-those arising from biases in the data and those arising from the algorithms. Firstly, data is often heterogeneous, generated by subgroups with their own characteristics and behaviors. Then a model learned on biased data may lead to unfair and inaccurate predictions [6, 7]. Secondly, for the algorithmic fairness, one should first define the notion of fairness to fight against discrimination and achieve fairness. However, the fact that no universal definition of fairness exists shows the difficulty of solving this problem. With different definitions, a variety of methods have been proposed that satisfy some of the fairness definitions or other new definitions depending on the application [8, 9, 10, 11, 12].

The goal of this paper is to extend the framework established in our work [2] to address the fairness and privacy issues simultaneously in the IAS design. The main observation is that the transformation map employed in [2] could not only be used for privacy protection but could also be used for fairness representation. However, there is a trade-off among data utility, fairness representation and privacy protection. By carefully designing a transformation map on the original data, the predictor will not observe the data directly, thereby reducing the bias and enhancing the privacy protection, but it will also reduce the inference accuracy. To properly address the trade-off between different goals, we formulate an optimization problem to find the optimal transformation map. To quantify the inference accuracy, we use mutual information between the transformed variable and the label. To guarantee the fairness, we measure the bias by mutual information between the transformed variable and the sensitive attribute. To determine the privacy protection, instead of using a specific privacy leakage measure, we follow our previous work [2] and apply a general privacy leakage metric defined by a continuous function f, where different choices of f lead to different

This work was supported in part by National Science Foundation under Grants CCF-1717943, CNS-1824553, CCF-1908258 and ECCS-2000415.

privacy measures. Thus, the trade-off problem can be solved through a maximization problem where the objective function is composed of the above-mentioned three terms. To solve the maximization problem, if we optimize over the space of the transformation map directly, the formulated problem is non-convex with multiple constraints. Through various transformations and variable augmentations, we notice that there are four dominating arguments with certain nice property. We then exploit this structure and design an algorithm to solve the optimization problem by iterating between dominating arguments until reaching convergence. Compared with solving the optimization problem using gradient ascent in the space of the transformation map directly, the proposed method has better performance in the convergence speed and solution quality.

2. PROBLEM FORMULATION

As shown in Fig.1, we consider an inference as service problem, in which one would like to infer the parameter $S \in \mathcal{S}$ of data $Y \in \mathcal{Y}$, in which $\mathcal{Y}$ has a finite alphabet. At the meantime, there is a sensitive attribute Z which contains sensitive information such as race, gender etc. Under this setup, instead of sending Y directly to the server, we will learn a transformation map from Y to $U \in \mathcal{U}$, and send U to the server. The server will use U to conduct the inference task. This transformation mapping serves two purposes: fair presentation and privacy protection. In order to mitigate the bias, we seek to find U that captures all the relevant information to predict S while not containing any information about the sensitive attribute Z. To preserve the privacy, we want U to disclose as little information about Y as possible. Here, $\mathcal{U}$ also has a finite alphabet and is allowed to be different from $\mathcal{Y}$. Without loss of generality, we will employ a randomized mapping and use $p(u|y)$ to denote the probability that data $Y = y$ will be mapped to $U = u$ and the whole mapping is denoted as $P_{U|Y}$. Furthermore, we use P_S to denote the prior distribution of S, $P_{Z|S}$ to denote the conditional distribution Z given S and $P_{Y|S}$ to denote the conditional distribution Y given S, while the lower-case letter p is used to denote the component-wise probability (e.g., $p(s), p(z|s), p(y|s)$ will be used in the sequel). Thus, Z, Y, U form a Markov chain, and S, Y, U form another Markov chain.

To measure the inference accuracy, note that the distributional difference between P_S and $P_{S|U}$ characterizes the information about S contained in U. As $I(S;U)$ is the averaged Kullback-Leibler (KL) divergence between P_S and $P_{S|U}$, we use it to measure the inference accuracy. We would like to make $I(S;U)$ as large as possible, which means that we would like retain as much information about the parameter of interest S in U so that the server can make a more accurate inference.

To measure the fairness, note that the distributional difference between P_Z and $P_{Z|U}$ characterizes the information about Z contained in U, which is related to the bias. Since the inference is based on U and $I(Z;U)$ is the averaged KL divergence between P_Z and $P_{Z|U}$, we use it to measure the bias. We would like to make $I(Z;U)$ as small as possible so that U contains as little information about the sensitive attribute Z as possible.

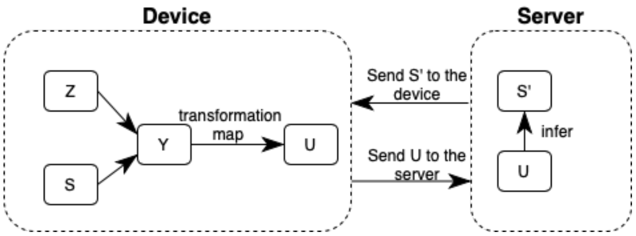


Fig. 1. Problem setup: Z is the sensitive attribute, S is the parameter of interest, Y is the data observation, U is the transformed variable after the transformation map and S' is the inferred result.

To measure the privacy leakage, we follow similar approach as in our recent work [2]. In particular, we choose a general form $\mathbb{E}_{Y,U}[d(y,u)]$, in which $d(y,u) = f(\frac{p(y)}{p(y|u)})$ and f is a continuous function defined on $(0,+\infty)$. We note that $\mathbb{E}_{Y,U}[d(y,u)] = \mathbb{E}_{Y,U}[f(\frac{p(y)}{p(y|u)})]$ measures the distributional distance between P_Y and $P_{Y|U}$, where P_Y is the prior distribution of Y and $P_{Y|U}$ is the posterior distribution of Y after observing U. Hence, the smaller the distance, the less information U can provide about Y and the better the privacy protection. This form is applicable for different privacy metrics by setting f in different forms [2].

Taking all these three conflicting goals into consideration, we aim to find the mapping $p(u|y)$ that solves the following optimization problem

$$\max_{P_{U|Y}} \ \mathcal{F}[P_{U|Y}] \triangleq I(S;U) - \beta \mathbb{E}_{Y,U}\left[f\left(\frac{p(u|y)}{p(u)}\right)\right] - \alpha I(Z;U), \tag{1}$$

$$\text{s.t. } \ p(u|y) \geq \epsilon, \forall y, u, \sum_{u} p(u|y) = 1, \forall y \in \mathcal{Y}. \tag{2}$$

Here, $\alpha \in (0,\infty)$ is a weight that indicates the relative importance of minimizing the bias and $\beta \in (0,\infty)$ is a weight that indicates the relative importance of maximizing the privacy protection.

The problem setting is an extension of our previous work [2], which investigates the privacy-accuracy trade-off in the IAS scenario. The concept of measuring the inference accuracy and the privacy leakage follows directly from [2]. However, by taking fairness issue into consideration, we have an additional term $\alpha I(Z;U)$ in the objective function, which is non-convex with respect to $P_{U|Y}$. Thus, the formulated optimization problem is much more complicated.

3. PROPOSED METHOD

In this section, we discuss how to solve the optimization problem defined in (1). As the objective function is a complicated non-convex function of $P_{U|Y}$, we only expect to find the local maximal point. First, we transform the maximization over single argument to an alternative maximization problem over multiple arguments. Then the Alternating Direction Method of Multipliers(ADMM) method is introduced to solve the subproblems.

From [2], we have that

$$I(S;U) = I(S;Y) - \sum_{u,y} p(y)p(u|y)D_{KL}[p(s|y) \parallel p(s|u)].$$

Then the objective function defined in (1) can be written as

$$\begin{aligned} \mathcal{F}[P_{U|Y}, P_U, P_{Z|U}, P_{S|U}] \quad &= \quad I(S;Y) + \beta \mathbb{E}_{Y,U}[d(y,u)] \\ - \sum_{u,y} p(y)p(u|y)&D_{KL}[p(s|y) \parallel p(s|u)] - \alpha I(Z;U). \end{aligned}$$

For consistency, we require the following equations to be satisfied simultaneously

$$p(u) = \sum_y p(u|y)p(y), \forall u, \tag{3}$$

$$p(z|u) = \frac{\sum_y p(u|y)p(z,y)}{p(u)}, \tag{4}$$

$$p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}. \tag{5}$$

By (4) and (5) , we require that $p(u) > 0, \forall u$.

By considering the objective function defined in (1) as a functional on $P_{U|Y}$, P_U, $P_{Z|U}$ and $P_{S|U}$, we have the following lemma.

Lemma 1. *Suppose that $f(\cdot)$ is a strictly convex function. Then for given $P_U, P_{Z|U}, P_{S|U}$, $\mathcal{F}[P_{U|Y}, P_U, P_{Z|U}, P_{S|U}]$ is concave in each $P_{U|y_i}, \forall y_i \in \mathcal{Y}$. Similarly, for given $P_{U|Y}, P_{Z|U}, P_{S|U}$, $\mathcal{F}[P_{U|Y}, P_U, P_{Z|U}, P_{S|U}]$ is concave in P_U. For given $P_{U|Y}, P_U, P_{S|U}$, $\mathcal{F}[P_{U|Y}, P_U, P_{Z|U}, P_{S|U}]$ is concave in $P_{Z|U}$. For given $P_{U|Y}, P_U, P_{Z|U}$, $\mathcal{F}[P_{U|Y}, P_U, P_{Z|U}, P_{S|U}]$ is concave in $P_{S|U}$.*

Then a natural approach to solve (2) with the requirements on the dominating arguments is to alternately iterate between $P_{U|Y}$, P_U, $P_{Z|U}$ and $P_{S|U}$ until reaching convergence. Following this insight, we rewrite (2) as an alternating optimization problem

$$\begin{aligned} &\max_{P_{S|U}} \max_{P_{Z|U}} \max_{P_U} \max_{P_{U|Y}} \mathcal{F}[P_{U|Y}, P_U, P_{Z|U}, P_{S|U}]. \\ &\text{s.t.} \quad p(u|y) \geq \epsilon, \forall y, u, \ \sum_u p(u|y) = 1, \forall y, \\ &\qquad\ \ p(u) > 0, \forall u, \ \sum_u p(u) = 1, (3), \\ &\qquad\ \ p(z|u) \geq 0, \forall u, z, \ \sum_z p(z|u) = 1, \forall u, (4), \\ &\qquad\ \ p(s|u) \geq 0, \forall u, s, \ \sum_s p(s|u) = 1, \forall u, (5). \end{aligned}$$

Under this formula, we will solve the maximization on $P_{S|U}$ first and derive an analytical result as a function of P_U, $P_{U|Y}$ and $P_{Z|U}$. Then consider the maximization on P_U, $P_{U|Y}$ and $P_{Z|U}$ for a given $P_{S|U}$.

For $P_{S|U}$, the maximization problem is

$$\begin{aligned} \max_{P_{S|U}} \quad & \mathcal{F}[P_{S|U}|P_{U|Y}, P_U, P_{Z|U}], \\ \text{s.t.} \quad & p(s|u) \geq 0, \forall u, s, \sum_s p(s|u) = 1, \forall u, (5). \end{aligned}$$

The solution can be easily derived as $p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}$, which satisfies all the constraints naturally.

Then we update $P_{Z|U}$ by the consistency equation (4). For the given $P_{S|U}$ and $P_{Z|U}$, we solve the optimization problem on $P_{U|Y}$ and P_U. Since it is a non-convex problem with multiple constraints, we apply ADMM to solve the problem. The augmented Lagrangian for the above problem is given by

$$\begin{aligned} &\mathcal{L}[P_{U|Y}, P_U, P_{S|U}, P_{Z|U}; \Lambda] \\ = &\mathcal{F}[P_{U|Y}, P_U|P_{S|U}, P_{Z|U}] + \sum_u \lambda(u)\delta(u) - \frac{\rho}{2}\sum_u \delta(u)^2, \end{aligned}$$

where Λ is a vector of size $|\mathcal{U}|$. Then the optimization problem on $P_{U|Y}$ and P_U can be solved by the following iterative procedure,

$$P_{U|y_i}^{t+1} = \arg\max_{P_{U|y_i}} \mathcal{L}[P_{U|y_i}, P_{U|Y^{(i-)}}^{t+1}, P_{U|Y^{(i+)}}^{t}, P_U^t; \Lambda^t], \tag{6}$$

$$P_U^{t+1} = \arg\max_{P_U} \mathcal{L}[P_{U|Y}^{t+1}, P_U; \Lambda^t], \tag{7}$$

$$\Lambda^{t+1} = \Lambda^t - \rho(P_U^{t+1} - (P_{U|Y}^{t+1})^T P_Y), \tag{8}$$

where $P_{U|Y^{(i-)}}$ denotes all rows before the i-th row in the matrix $P_{U|Y}$ and $P_{U|Y^{(i+)}}$ denotes all rows after the i-th row.

After solving two sub-problems on $P_{U|Y}$ and P_U respectively, we update the value of Λ. The algorithm is given in Algorithm 1.

4. NUMERICAL RESULTS

In this section, we give numerical examples to show that our proposed formulation can mitigate the bias towards certain

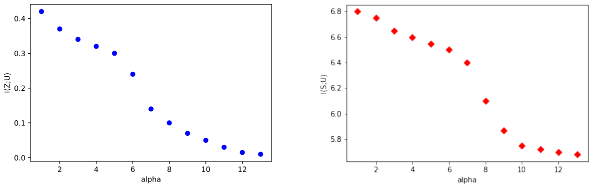
Algorithm 1 Design the optimal transformation map

```
Input:
Prior distribution $P_S, P_Z$ and conditional distribution $P_{Y|S,Z}$.
Trade-off parameter $\alpha, \beta$.
Converge parameter $\eta, \eta_p$.
Output:
A mapping $P_{U|Y}$ from $Y \in \mathcal{Y}$ to $U \in \mathcal{U}$.
Initialization:
Randomly initiate $P_{U|Y}$ and calculate $P_U, P_{Z|U}, P_{S|U}$ by (3), (4) and (5).
$j = 1$.
while $\left\| P_{S|U}^{(j)} - P_{S|U}^{(j-1)} \right\|_F > \eta$ do
    $P_U^{(j),1} = P_U^{(j-1)}$.
    $P_{U|Y}^{(j),1} = P_{U|Y}^{(j-1)}$.
    $t = 1$.
    while $t = 1$ or $\left\| P_U^{(j),t} - P_U^{(j),t-1} \right\|_{\ell_1} > \eta_p$ do
        Update $P_{U|y_i}$ by solving (6).
        Update $P_U$ by solving (7).
        Update $\Lambda$ by (8).
        $t = t + 1$.
    Update $P_{Z|U}^{(j)}$ by (4).
    Update $P_{S|U}^{(j)}$ by (5).
    $j = j + 1$.
return $P_{U|Y}$
```

groups. Moreover, for the proposed optimization problem, our proposed algorithm converges much faster than GA, and the solution found has better quality than the one found by GA.

To simplify the simulation, we suppose that $Z \in \{0, 1\}$, which could represent sensitive information such as gender, race, etc. We set the prior distributions $\boldsymbol{p}_z = \{\frac{1}{4}, \frac{3}{4}\}$ and $\boldsymbol{p}_s = \{\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\}$. Let $|\mathcal{Y}| = 9, |\mathcal{U}| = 11$. The conditional distributions $P_{Y|S}(y|s, Z = 0)$ and $P_{Y|S}(y|s, Z = 1)$ are randomly generated. The initial map $P_{U|Y}$ is obtained by selecting uniformly distributed random numbers and normalizing them. By setting f as $f(x) = x \log \frac{2x}{x+1} + \log \frac{2}{x+1}$, we use Jensen-Shannon divergence as the privacy leakage measure. Then we will perform both Algorithm 1 and GA to find the transformation map.

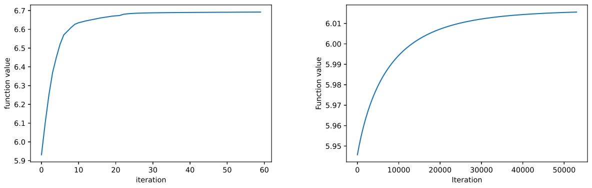
Firstly, we explore the relationship between fairness trade-off parameter α and the degree of fairness. Set the privacy trade-off parameter $\beta = 7$. Then we randomly initialize $P_{U|Y}$ and run the algorithm until it terminates for different αs. The stopping criterion is $||P_{U|Y}^{t+1} - P_{U|Y}^{t}||_F < 10^{-4}$. We repeat this procedure 300 times for each α. As shown in Fig. 2(a), we notice that the bias measure $I(Z; U)$ is decreasing as α increases, indicating that the transformed variable U provides less information about the sensitive attribute Z and thus the predictor will discriminate less against certain groups.



(a) $I(Z, U)$ v.s. α

(b) $I(S; U)$ v.s. α

Fig. 2. Effects of α



(a) Function value of Algorithm 1

(b) Function value of GA

Fig. 3. Function value v.s. iteration

Secondly, we also explore the relationship between α and the information accuracy. As shown in Fig. 2(b), the information accuracy term $I(S; U)$ is decreases as α increases, indicating that the predictive ability becomes weaker. However, Fig. 2(b) also shows that the reduction of $I(S; U)$ is not very large, which implies that the model still has good predictive ability under the condition of less discrimination.

Thirdly, we investigate the convergence speed of the proposed algorithm. Fig. 3(a) illustrates the relationship between objective function values and iteration number. This figure shows that the objective function value monotonically increases and converges as the iterative process progresses. For comparison purpose, we also plot the corresponding figure for GA in Fig. 3(b). From these figures, we can see that Algorithm 1 converges within 30 iterations. Nevertheless, for GA, it is difficult to determine a proper step size and the optimal function value found by GA is always smaller than the value found by Algorithm 1.

5. CONCLUSION

We have explored the utility, fairness and privacy trade-off in IAS scenarios under sensitive environments. We have formulated an optimization problem to find the desirable transformation map. We have transformed the formulated non-convex optimization problem and designed an iterative method to solve it. Moreover, we have provided numerical results showing that the proposed method can mitigate the bias and has better performance than GA in the convergence speed, solution quality and algorithm stability.

6. REFERENCES

[1] A. Gujarati, S. Elnikety, Y. He, K. McKinley, and B. Brandenburg, “Swayam: Distributed autoscaling to meet SLAs of machine learning inference services with resource efficiency,” in *Proc. ACM/IFIP/USENIX Middleware Conference*, (Las Vegas, NV), pp. 109–120, Dec. 2017.

[2] Y. Jin and L. Lai, “Privacy-accuracy trade-off of inference as service,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, (Toronto, Canada), pp. 2645–2649, Jun. 2021.

[3] M. Samragh, H. Hosseini, A. Triastcyn, K. Azarian, J. Soriaga, and F. Koushanfar, “Unsupervised information obfuscation for split inference of neural networks,” *arXiv preprint arXiv:2104.11413*, Apr. 2021.

[4] E. Erdemir, P. L. Dragotti, and D. Gündüz, “Active privacy-utility trade-off against a hypothesis testing adversary,” in *proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2660–2664, Jun. 2021.

[5] S. Corbett-Davies and S. Goel, “The measure and mismeasure of fairness: A critical review of fair machine learning,” *arXiv preprint arXiv:1808.00023*, Aug. 2018.

[6] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, “Social data: Biases, methodological pitfalls, and ethical boundaries,” *Frontiers in Big Data*, vol. 2, p. 13, Jul. 2019.

[7] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, “The dataset nutrition label: A framework to drive higher data quality standards,” *arXiv preprint arXiv:1805.03677*, May 2018.

[8] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in *Proc. International conference on machine learning*, (Atlanta, Georgia), pp. 325–333, PMLR, Jun. 2013.

[9] Y. Bechavod and K. Ligett, “Penalizing unfairness in binary classification,” *arXiv preprint arXiv:1707.00044*, Jun. 2017.

[10] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil, “Empirical risk minimization under fairness constraints,” in *Proc. International Conference on Neural Information Processing Systems*, (Siem Reap, Cambodia), pp. 2796–2806, Dec. 2018.

[11] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, “Decoupled classifiers for group-fair and efficient machine learning,” in *Proc. Conference on Fairness, Accountability and Transparency*, (New York, NY), pp. 119–133, PMLR, Feb. 2018.

[12] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Proc. International Conference on Neural Information Processing Systems*, (Barcelona Spain), pp. 3323–3331, Dec. 2016.