Defensive Collaborative Learning: Protecting Objective Privacy in Data Sharing

Extended Abstract

Cynthia Huang
University of Waterloo, Waterloo, Canada
Vector Institute, Toronto, Canada
hhuang@uwaterloo.ca

Pascal Poupart
University of Waterloo, Waterloo, Canada
Vector Institute, Toronto, Canada
ppoupart@uwaterloo.ca

ABSTRACT

In collaborative machine learning, protecting proprietary information and safeguarding competitive advantages are crucial for participating organizations. This necessitates the development of algorithms that target a general notion of privacy defined by the data owner: objective privacy. In this paper, we formalize the idea of objective privacy as the protection of private value propositions characterized by predictive functions. We propose Defensive Collaborative Learning (DCL), where participants share data collaboratively while safeguarding their objective privacy. Formulating a min-max optimization problem that trades off utility and privacy protection, we propose algorithms that leverage mutual information backpropagation in both decentralized and centralized settings. Empirical studies show that the proposed algorithms protect objective privacy while enabling data sharing.

KEYWORDS

Collaborative Learning; Machine Learning Privacy

ACM Reference Format:
Cynthia Huang and Pascal Poupart. 2023. Defensive Collaborative Learning: Protecting Objective Privacy in Data Sharing: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

When organizations participate in collaborative machine learning by sharing data, they are motivated to protect proprietary information underlying their business value propositions, particularly in competitive industries like finance. In the case of a bank exchanging customer data with an e-commerce company to make better data-driven decisions, such as whether to extend credit to a customer, there is a concern that the e-commerce company will also offer similar credit products. Banking data could provide valuable information about credit profiles and help them compete.

This calls for a notion of privacy that captures the objective defined by the data owner. Consider a dataset $D = \{(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}\}$ and a predictive function $g : \mathcal{X} \to \mathcal{Y}$ that characterizes an objective. A data-sharing algorithm $\mathcal{M} : \mathcal{X} \to \mathcal{Z}$ is *objective private* when it produces a representation $\mathcal{Z}$ of the data $\mathcal{X}$ for sharing such that if any participant were to perform an objective privacy attack (i.e., infer g), the attack's performance gained from $\mathcal{Z}$ being shared is minimized. In the previous example, the bank's objective could be its proprietary function in determining if a customer would default.

This paper considers *Defensive Collaborative Learning* (DCL): a vertical data-sharing setting where participants share data representations while safeguarding knowledge pertaining to their self-defined objectives. The objectives can be customized data privacy functions for individuals or data-driven competitive advantages in multi-organizational learning. Advances in DCL can facilitate customized data privacy protection, a sophisticated data market, and trustworthy multi-organizational learning.

2 DEFENSIVE COLLABORATIVE LEARNING

2.1 Problem Formalization

Consider the following data-sharing collaborative learning scenario: there are n participants, each with their own dataset $\mathcal{X}_i$. Here the dataset is vertically split; that is, there is a unique identifier that enables data linkage between $\mathcal{X}_i$ and $\mathcal{X}_j$ to produce a rich representation for every data record. Let $\mathcal{X} = \mathcal{X}_1 \otimes \mathcal{X}_2 \cdots \otimes \mathcal{X}_n$ denote the aggregated data containing all the features where $\otimes$ denotes a join operation based on the unique identifier. Additionally, each participant has an objective that she would like to protect. We define the *objective* for participant i as the best attainable *predictive function* g_i^* with range $\mathcal{Y}_i$, representing the space of output targets.

In DCL, participants share data representations to minimize information loss about the original data while protecting their objective privacy. We assume that participant i utilizes an encoder $h_i : \mathcal{X}_i \to \mathcal{Z}_i$ to embed $\mathbf{x}_i$ into $\mathbf{z}_i$ (i.e., $h_i(\mathbf{x}_i) = \mathbf{z}_i$). Then we formalize the problem as follows:

$$\min_{h_i, k_i} \mathbf{E}[L_i^{rep}(\mathbf{X}_i, k_i(h_i(\mathbf{X}_i)))]$$

$$\text{s.t.} \min_{f_j} \mathbf{E}[L_i^{pred}(g_i^*(\mathbf{X}), f_j(\mathbf{X}_j, h_i(\mathbf{X}_i)))] \geq t, \forall j \neq i$$

Here, $k_i : \mathcal{Z}_i \to \mathcal{X}_i$ denotes the decoder that reconstructs the inputs from the embeddings. We use L_i^{rep} to denote the reconstruction loss of $\mathbf{X}_i$ with mean squared error (MSE) as the specific loss function. f_j refers to a predictive function for the output targets $g_i^*(\mathbf{X})$ based on $\mathbf{X}_j$ and $h_i(\mathbf{X}_i)$. L_i^{pred} denotes the specific prediction loss. Finally, t denotes a threshold for the expected prediction loss. When $t = \min_{f_j} \mathbf{E}[L_i^{pred}(g_i^*(\mathbf{X}), f_j(\mathbf{X}_j))]$, the shared representation provides no additional valuable information for predicting the target. In practice, participant i approximates the objective function g_i^* and uses the estimated values in the optimization, for example, by training a predictor $\hat{g}_i : \mathcal{X}_i \to \mathcal{Y}_i$ based on $\mathcal{X}_i$ only.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

2.2 Algorithms

Based on information theory, we approximate the constraint as a minimization of conditional mutual information between $\mathbf{X}_i$ and $\mathbf{Z}_i$ given $\mathbf{X}_j$ [3]. By leveraging ideas from mutual information backpropagation [4, 5], we propose the following algorithms that use a mutual information neural estimator [1].

2.2.1 Decentralized Defensive Collaborative Learning via Mutual Information Backpropagation (DDCL-MI).

In decentralized data sharing, each participant generates representations independently. Thus, $\mathbf{X}_j$ is not available for participant i and we minimize the mutual information between $\mathbf{X}_i$ and $\mathbf{Z}_i$ instead. We adopt a general-purpose mutual information neural estimator proposed by [1], which maximizes the following loss function to learn $t^\phi : \mathcal{Z}_i \times \mathcal{Y}_i \to \mathbf{R}$ where $\mathbf{z}_i = h_i(\mathbf{x}_i) \in \mathcal{Z}_i$:

$$L_t^{mi} = \mathop{\mathbf{E}}_{(\mathbf{z}_i,\mathbf{y}_i)\sim P_{(Z_i,Y_i)}}[t^\phi(\mathbf{z}_i,\mathbf{y}_i)] - \log \mathop{\mathbf{E}}_{\mathbf{z}_i\sim P_{Z_i},\hat{\mathbf{y}}_i\sim P_{Y_i}}[\exp t^\phi(\mathbf{z}_i,\hat{\mathbf{y}}_i)]$$

Once maximized, L_t^{mi} is used as an estimate for the mutual information. Therefore, DDCL-MI optimizes the following objective:

$$\min_{h_i,k_i}\max_t \mathbf{E}[L_i^{rep}(\mathbf{X}_i, k_i(h_i(\mathbf{X}_i)))] + \alpha L_t^{mi}$$

2.2.2 Centralized Defensive Collaborative Learning via Conditional Mutual Information Backpropagation (CDCL-CMI).

In centralized data sharing, a trusted central server generates representations with knowledge of all participants' data and objectives. The server aims to maximize information sharing and protect objective privacy for all participants. With access to $\mathbf{X}_j$, the central server minimizes the conditional mutual information between $\mathbf{Z}_i$ and $\mathbf{Y}_i$ given $\mathbf{X}_j$, i.e. $MI(\mathbf{Z}_i, \mathbf{Y}_i|\mathbf{X}_j)$. By directly estimating the conditional mutual information with CMIGAN [4], learning instability was induced due to the complexity of the two levels of min-max optimization. Noticing that $MI(\mathbf{Y}_i, \mathbf{Z}_i|\mathbf{X}_j) = MI(\mathbf{Y}_i, (\mathbf{Z}_i, \mathbf{X}_j)) - MI(\mathbf{Y}_i, \mathbf{X}_j)$ and $MI(\mathbf{Y}_i, \mathbf{X}_j)$ does not depend on the encoder, we solve for the following optimization problem instead:

$$\min_{h_i,k_i} \mathbf{E}[L_i^{rep}(\mathbf{X}_i, k_i(h_i(\mathbf{X}_i)))] + \alpha MI(\mathbf{Y}_i, (\mathbf{Z}_i, \mathbf{X}_j))$$

This enables us to use a mutual information neural estimator, similar to DDCL-MI. We refer to $MI(\mathbf{Y}_i, (h_i(\mathbf{X}_i), \mathbf{X}_j))$ as the joint mutual information (Joint-MI). Again, we use neural network $t^\phi : \mathcal{Z}_i \times \mathcal{X}_j \times \mathcal{Y}_i \to \mathbf{R}$ with weight ϕ which maximizes the following loss function via backpropagation:

$$L_t^{cmi} = \mathop{\mathbf{E}}_{(\mathbf{z}_i,\mathbf{x}_j,\mathbf{y}_i)\sim P_{(Z_i,X_j,Y_i)}}[t^\phi(\mathbf{z}_i,\mathbf{x}_j,\mathbf{y}_i)] - \log \mathop{\mathbf{E}}_{\mathbf{z}_i,\mathbf{x}_j\sim P_{(Z_i,X_j)},\hat{\mathbf{y}}_i\sim P_{Y_i}}[\exp t^\phi(\mathbf{z}_i,\mathbf{x}_j,\hat{\mathbf{y}}_i)]$$

Once maximized, L_t^{cmi} can be used as an estimate of Joint-MI. Therefore, CDCL-CMI optimizes for the following objective:

$$\min_{h_i,k_i}\max_t \mathbf{E}[L_i^{rep}(\mathbf{X}_i, k_i(h_i(\mathbf{X}_i)))] + \alpha L_t^{cmi}$$

3 EXPERIMENTAL STUDY

We evaluate the algorithms on a curated two-party collaborative learning settings based on the Adult UCI dataset [2]. After feature transformation, the dataset size is 48842×66. The features are vertically partitioned between two participants. Assuming a two-layer neural network as the underlying predictive function, the objectives are generated as a vector of dimensionality four. For categorical objectives, we use the softmax activation function and for continuous objectives, we use the identity function.

To evaluate the effectiveness of data sharing, we consider the performance of an *objective predictor*, which is trained based on each participant's own data as well as the shared representations. Reconstruction error is also reported in MSE as an alternative measure for data sharing utility. As a measure of objective privacy protection, we examine the performance of an *objective inference attacker*, which predicts the other participants' target outputs. As baseline, we also consider the following settings: (1) *ML*: participants do not share data; (2) *CL-AE*: participants share the embeddings from the regular autoencoder.

Table 1: Performance of Algorithms - Adult UCI

Data Type	Algorithm	Rep MSE ↓	Obj Acc/R^2 ↑	Attack Acc/R^2 ↓
Continuous	ML	-	72.46 ± 0.27	37.25 ± 0.22 *
	DDCL-MI	0.48 ± 0.016	90.62 ± 1.98 *	79.91 ± 5.52 *
	CDCL-CMI	1.14 ± 0.43	88.91 ± 2.04 *	62.71 ± 5.45 *
	CL-AE	0.09 ± 0.028	97.71 ± 0.66 *	95.48 ± 0.60
Discrete	ML	-	81.01 ± 3.04	73.67 ± 4.49 *
	DDCL-MI	0.36 ± 0.061	94.11 ± 1.91 *	86.17 ± 3.04 *
	CDCL-CMI	1.04 ± 0.403	84.25 ± 3.67 *	79.23 ± 2.63 *
	CL-AE	0.08 ± 0.029	95.98 ± 1.93 *	93.77 ± 1.85

Notes: Accuracy and R^2 is expressed in % and MSE is expressed in units of 1e-2. For objective accuracy, Wilcoxon signed rank test p-value compares algorithms with ML. For attack accuracy, Wilcoxon signed rank test p-value compares algorithms with CL-AE. * indicates p<0.01.

As shown in the table above, DCL algorithms, with a higher reconstruction error than CL-AE, deliver data sharing benefits. Compared to ML, they offer better objective prediction performance. In terms of objective privacy protection, DCL algorithms perform much better than CL-AE, but not as well as ML. As compared to DDCL-CMI, CDCL-CMI offers better privacy protection at the expense of lower data sharing utility. Overall, DCL algorithms preserve data sharing utility while protecting objective privacy.

4 CONCLUSIONS AND FUTURE DIRECTIONS

Motivated by real-life collaborative machine learning scenarios, we propose Defensive Collaborative Learning that ensures objective privacy. By introducing the min-max optimization formulation for DCL, we formalize objective privacy under potential objective privacy attacks. Furthermore, we propose and examine mutual information backpropagation algorithms in centralized and decentralized settings. Based on experiments, these algorithms provide data sharing benefits while limiting objective privacy leakage.

REFERENCES

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholm, Sweden, 531–540. https://proceedings.mlr.press/v80/belghazi18a.html

[2] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[3] Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. 2014. From the Information Bottleneck to the Privacy Funnel. In *2014 IEEE Information Theory Workshop (ITW 2014)*. IEEE, Hobart, Tasmania, Australia, 501–505. https://doi.org/10.1109/ITW.2014.6970882

[4] Arnab Kumar Mondal, Arnab Bhattacharjee, Sudipto Mukherjee, Himanshu Asnani, Sreeram Kannan, and Prathosh A. P. 2020. C-MI-GAN : Estimation of Conditional Mutual Information using MinMax formulation. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, Ryan P. Adams and Vibhav Gogate (Eds.). AUAI Press, online, 358. http://www.auai.org/uai2020/proceedings/358_main_paper.pdf

[5] Ruggero Ragonesi, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. 2021. Learning Unbiased Representations via Mutual Information Backpropagation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, online, 2723–2732. https://doi.org/10.1109/CVPRW53098.2021.00307