Privacy-preserving entity resolution and logistic regression on encrypted data

Mentari Djatmiko [1] Stephen Hardy [1] Wilko Henecka [1] Hamish Ivey-Law [1] Maximilian Ott [1] Giorgio Patrini [1]
Guillaume Smith [1] Brian Thorne [1] Dongyao Wu [1]

We consider a scenario of two data providers, A and B, each of whom manage a dataset of private information consisting of two different feature sets related to common customers/entities. They *jointly* aim to learn a linear model using stochastic gradient algorithms like SGD/SAG (Schmidt et al., 2013). The setting is federated learning (Konecný et al., 2016), where data is kept locally and a shared model is learned on top of local computation. Notice that, in contrast with the large majority of work on distributed learning, in our scenario data is split *vertically*, *i.e.* by features. We also assume that only A knows the target variable.

We propose a secure system solving the problem in two phases: privacy-preserving entity resolution and logistic regression over encrypted data. With the aid of a coordinator, C, we design a three-party protocol that is secure under the honest-but-curious adversary model. *Our system allows A and B to learn a classifier collaboratively, without either exposing their data in the clear or even sharing which entities they have in common.*

Privacy-preserving entity resolution When the dataset is vertically partitioned across multiple organisations the problem arises of how to identify the corresponding entities, namely *entity resolution* (Christen, 2012). Entity resolution is usually done on identifying features such as name, address, *etc*. We perform privacy-preserving entity resolution using anonymous linkage codes, which map entity information onto a code from which it is impossible to reconstruct any entity data. We use the *cryptographic longterm key* (CLK) (Schnell et al., 2011) anonymous linkage code, which provides both privacy and error tolerance. The CLKs are used in a fuzzy comparison function which allows us to perform an *inner join* on the two datasets.

Parties A and B create CLKs for each entry in their datasets and sent them to C, which performs the entity resolution. The protocol results in two *permutations*, one for each data provider, and a *mask*. The permutations describe how A and B should rearrange their dataset so as to be consistent with each other and the mask. The mask specifies whether a row corresponds to a record available in both datasets, thus a record which will be used for learning; it also implicitly excludes records that are not matched accross A and B. The mask itself is only sent to data providers in encrypted form to prevent revealing the common entities. For simplicity, we omit mention of the permutations and mask in what follows.

Logistic regression on encrypted data Learning is performed on data encrypted with the Paillier *partially homomorphic encryption scheme* (Paillier, 1999), an asymmetric scheme which permits both adding encrypted values and scaling encrypted values by unencrypted ones. These properties allow us to implement most of the linear algebra necessary for gradient descent optimization *on encrypted data*. Only C possesses the private key.

We approximate the logistic loss and its gradient via *Taylor expansion* around 0 which results in polynomials that A and B can evaluate collaboratively and securely by only transmitting intermediate values that are encrypted with the Paillier scheme. Experiments have shown that we can match the accuracy of exact logistic loss using a merely second-order Taylor approximation to the loss (hence linear approximation to the gradient) at the price of rescaling features into the interval $[-1, 1]$ and of applying L_1/L_2 regularization.

Party C orchestrates the optimization algorithm, taking care of the stochastic learning parameters (regularization, learning rate, momentum, etc.), triggering gradient computations by A and B, and using the logistic loss on hold-out data to determine when to stop training so as to avoid overfitting.

We have proven the practicality of our system in commercial deployments. Our system is capable of scaling to millions of records with hundreds of features.

References

Christen, P. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

Konecný, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527,

[1] Data61 CSIRO, Sydney, Australia. Correspondence to: Giorgio Patrini <giorgio.patrini@data61.csiro.au>.

2016. URL http://arxiv.org/abs/1610.02527.

Paillier, P. Public-key cryptosystems based on composite degree residuosity classes. In *International Conference on the Theory and Applications of Cryptographic Techniques*, 1999.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 2013.

Schnell, R., Bachteler, T., and Reiher, J. A novel error-tolerant anonymous linking code. Technical report, Paper No. WP-GRLC-2011-02, German Record Linkage Center Working Paper Series, 2011.