Towards Scalable, Efficient and Privacy Preserving Machine Learning

Rania Talbi, Sara Bouchenak

To cite this version:

HAL Id: hal-01956155
https://hal.archives-ouvertes.fr/hal-01956155
Submitted on 14 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Minimize the computational costs incurred by privacy preservation.

Provide an end-to-end privacy preserving outsourced data classification service.

Enable a set of mutually untrusted data owners to have a global vision on the union of their data without breaching the privacy of each one of them.

Enable dynamic data model updates when new training data samples are available.

Related work

- Clustering (1)
- Classification (2)
- Association Rule Mining (3)

Different Privacy-preservation objectives

- ML output protection (2)
- Original data protection (1)
- Non-cryptographic techniques (PP-Data Publishing techniques) (5)

Privacy Preservation techniques

- Distributed (4)
- Outsourced (5)

Design principles

- Decent privacy and utility levels
- Efficient runtime
- Entirely outsourced ML computations over encrypted data

Different ML algorithms

- Cryptographic based protection (data model, training data, classification queries and responses)
- Partial homomorphic encryption (PHE ) based building blocks
- Combine PHE with cryptographic blinding (DTPKC cryptosystem [6])

We implemented the VFDT incremental decision tree learning algorithm (7)

Naive approach: a combination of low level PP-building blocks

1st optimization : use inline building blocks

2nd optimization : Parallel computing

In our work,

- We used a synthetic dataset for fraud detection in a B2B network.
- This dataset contains 1000 bank transactions with 9 attributes each.
- We compare our work to the Ciphermed framework [8].

References

3. L. Liu et al.: Privacy-Preserving Mining of Association Rule on Outsourced Data from Multiple Parties. NDSS2018: 413-431
7. M. Domingo et al.: Mining high-speed data streams. ICIS 2000: 71-80
8. R. Bost et al.: Machine Learning Classification over Encrypted Data. NDSS 2015