Accuracy and Privacy Evaluations of Collaborative Data Analysis

Akira Imakura (University of Tsukuba), e-mail: imakura@cs.tsukuba.ac.jp

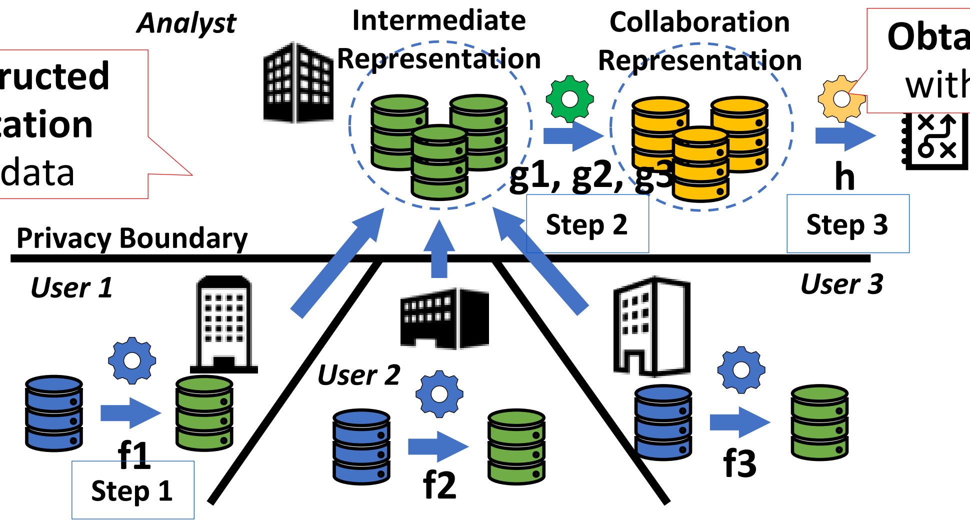
Co-authors: Anna Bogdanova, Takaya Yamazoe, Kazumasa Omote, Tetsuya Sakurai (University of Tsukuba)

Purpose

We analyze accuracy and privacy of collaborative data analysis (CDA), which shares dimensionality-reduced intermediate representation instead of the original data to obtain high-quality prediction for distributed data.

Share individually constructed intermediate representation
Instead of the original data

Analyze how to preserve the privacy of the original data



Obtain high-quality prediction
without sharing original data

Analyze how to guarantee the accuracy of CDA

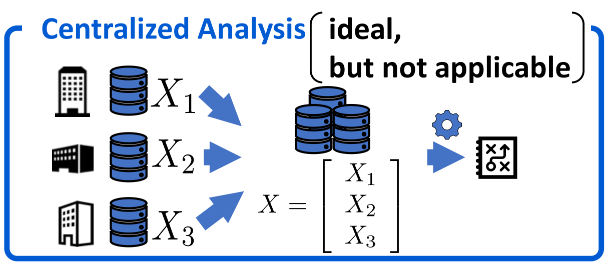
Contributions

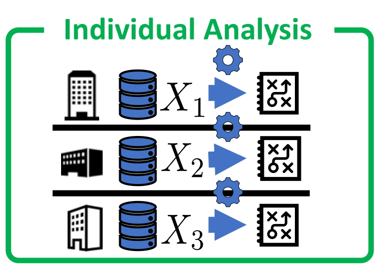
- ***For accuracy***: We provide ***a sufficient condition for equivalence*** of CDA and the centralized analysis with dimensionality reduction.
- ***For privacy***: We prove that, in CDA, the privacy of data is preserved by ***a double secureness***

Collaborative Data Analysis (CDA) [Imakura and Sakurai,2020]

Distributed data analysis

- ✓ In many applications (e.g., medicine, finance, and manufacturing), sharing original data for analysis may be difficult due to privacy and confidentiality requirements.





- ✓ Distributed data analysis aims to obtain high-quality prediction without sharing the original datasets

Step 1: Intermediate Representation (IR)

- ✓ Share individually constructed dimensionality-reduced intermediate representations (IR)
- ✓ ***Each party can use an individual function for IR***
  → Shared IR cannot be analyzed as one dataset

Step 2: Collaboration Representation (CR)

- ✓ Transform IR to an incorporable form named ***collaboration representation (CR)*** such that
  -- CR approximately match for the same "anchor data"

Step 3: Analysis

- ✓ Analyze CR as one dataset

Accuracy Analysis

Theoretical result

- ✓ ***Assumption:*** The functions for IR are linear, i.e.,

$$\widetilde{X}_i = f_i(X_i) = X_i F_i$$

Intermediate representation (IR) — Linear function (matrix) — Original data

- ✓ ***Theorem: A sufficient cond. for equivalence***

Sufficient cond. for equivalence

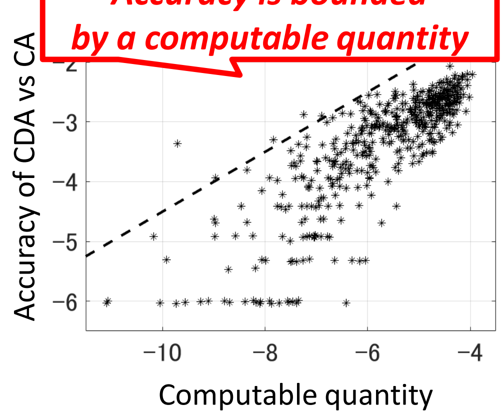
$$\mathcal{R}(F_1) = \cdots = \mathcal{R}(F_d)$$
$$\Downarrow$$

CDA = CA with dimensionality reduction

Numerical result

Accuracy is bounded by a computable quantity

- ✓ Numerical result of 500 trials for MNIST
- ✓ ***X-axis:*** log10 of computable quantity, which indicates approx. of the sufficient cond.
- ✓ ***Y-axis:*** log10 of 1 - NMI of CDA vs CA



Privacy Analysis

Protocol of CDA

- ✓ CDA is operated by ***users*** and ***analyst***.
  -- ***Users*** have the original data and construct IR, and share IR to analyst. (***The function for IR is not shared***)
  -- ***Analyst*** constructs CR and analyze them as one dataset

Theoretical result

- ✓ ***Target situation:***
  -- Each user want to protect the original data itself.
  -- ***Do not consider protecting statistics (e.g., average)***
- ✓ ***Theorem: A double secureness*** of CDA

In CDA, the privacy of the original data is preserved by ***a double secureness*** against insider and external attacks as

- ✓ No one can have the private data of others ***because each function is private*** under the protocol.
- ✓ Even if the function is stolen, the private data is still protected ***regarding ε-DR privacy***.

(ε-DR privacy means the original data cannot be recovered exactly from dimensionality reduced data even using the function.)