University of Glasgow

School of Computing Science
Knowledge & Data Engineering Systems

Self-Supervised Federated Learning over Relevant Heterogeneous Data

Tahani Aladwani, Christos Anagnostopoulos

Knowledge & Data Engineering Systems
School of Computing Science
University of Glasgow

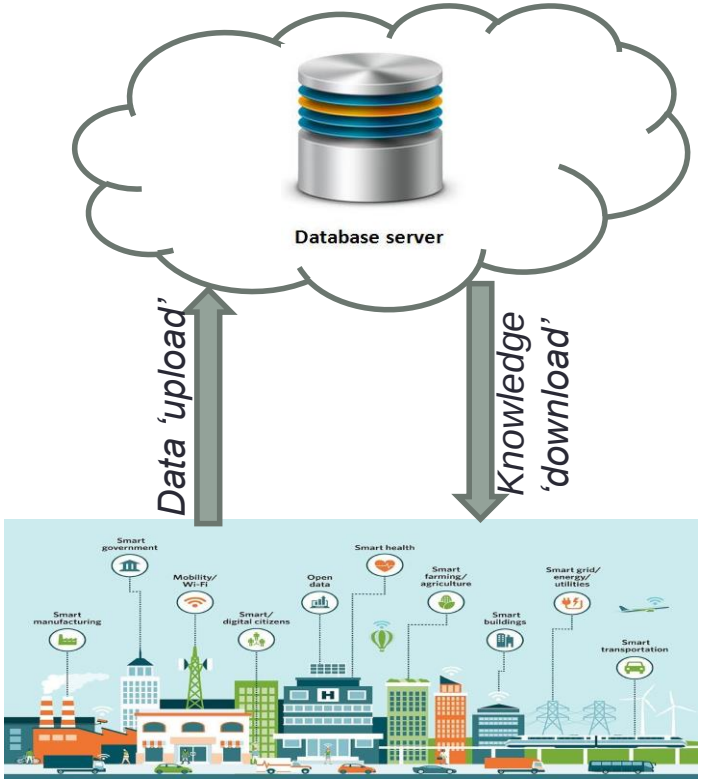
ADSAI Conference 2024 by The University of Manchester, 3-4 June 2024.

Introduction

Machine & Deep Learning (ML/DL) models are typically trained using centralized data.

However, bringing **all** data into a centralized server is no longer practical:

- **Violation of data privacy**
- **Communication burden due to data transfer**



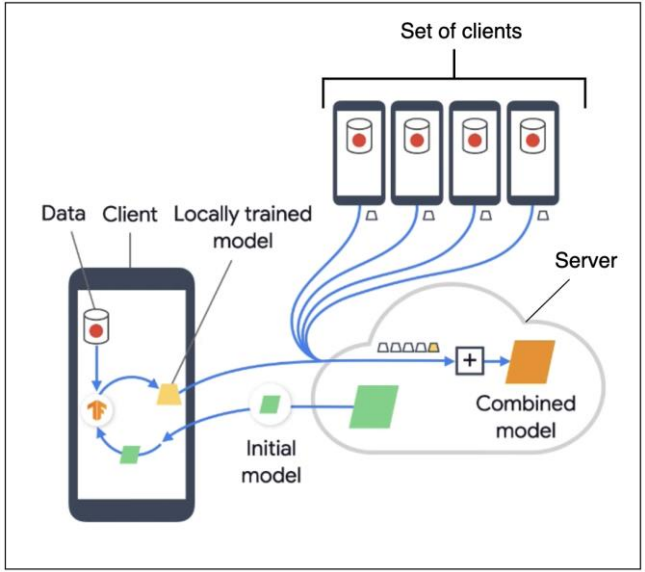
Distributed ML Model Training

Distributed Learning facilitates access to distributed data by training a ML model over *disjoint* data spaces by leveraging **nodes' local data and computational resources**.

Aim: Train a ML model efficiently requires training over a set of nodes, a.k.a., *participants*.

However, not all participants play the same role.
This is determined by:

- **Amount** of data and degree of heterogeneity.
- **Quality** of the data in each participant.
- **Lack of ground truth labels** (samples being unlabeled or partially labeled).



Problem Fundamentals

A typical FL system assumes that the labels on clients' datasets $n_i \in N$ and server's dataset D_S are identical. Both have the ground truth labels in the form of:

$\{\boldsymbol{X_i, y_i}\}$, X_i is input and y_i output.

Main label issues

- Possibility that a client $n_i \in N$ exhibits different label distribution y compared to server S, $P(y_i) \neq P(y_S)$.
- Each local dataset D_i may contain labels in {∅, 1, 2, . . . , L}, where ∅ represents samples without labels.
- Even if y_i and y_S have the same number of labels and utilize the same labelling mechanism, the labels cannot be considered entirely trustworthy.
  - **The ground truth labels are exclusively available at server S .**

We aim to solve these issues across clients in FL with a **Multi Purpose Semi Supervised Federated Learning paradigm (MP-SSFL).**

MP-SSFL Paradigm

Phase 1: Supervised learning: the server S builds a model based on its labeled data (D_S) to obtain: (X_S , y_S) $\rightarrow f(X_s, \theta_S)$. Then, distribute $f(., \theta_S)$ to each $n_i \in N$ in order to label their own data according to it.

Phase 2: Clients pseudo-labeling: Each client $n_i \in N$ uses $f(., \theta_S)$ to predict a pseudo-label $\hat{y}_i$ for each local sample $x_i \in n_i$.

- $\hat{y}_i$:the class with the highest predicted probability, i.e.,

$$\hat{y}_i = \arg \max(g\theta_S(x_i))_c$$

- $g\theta_S(x_i))$ represents the prediction probability of class *c* for input x_i.

MP-SSFL Paradigm

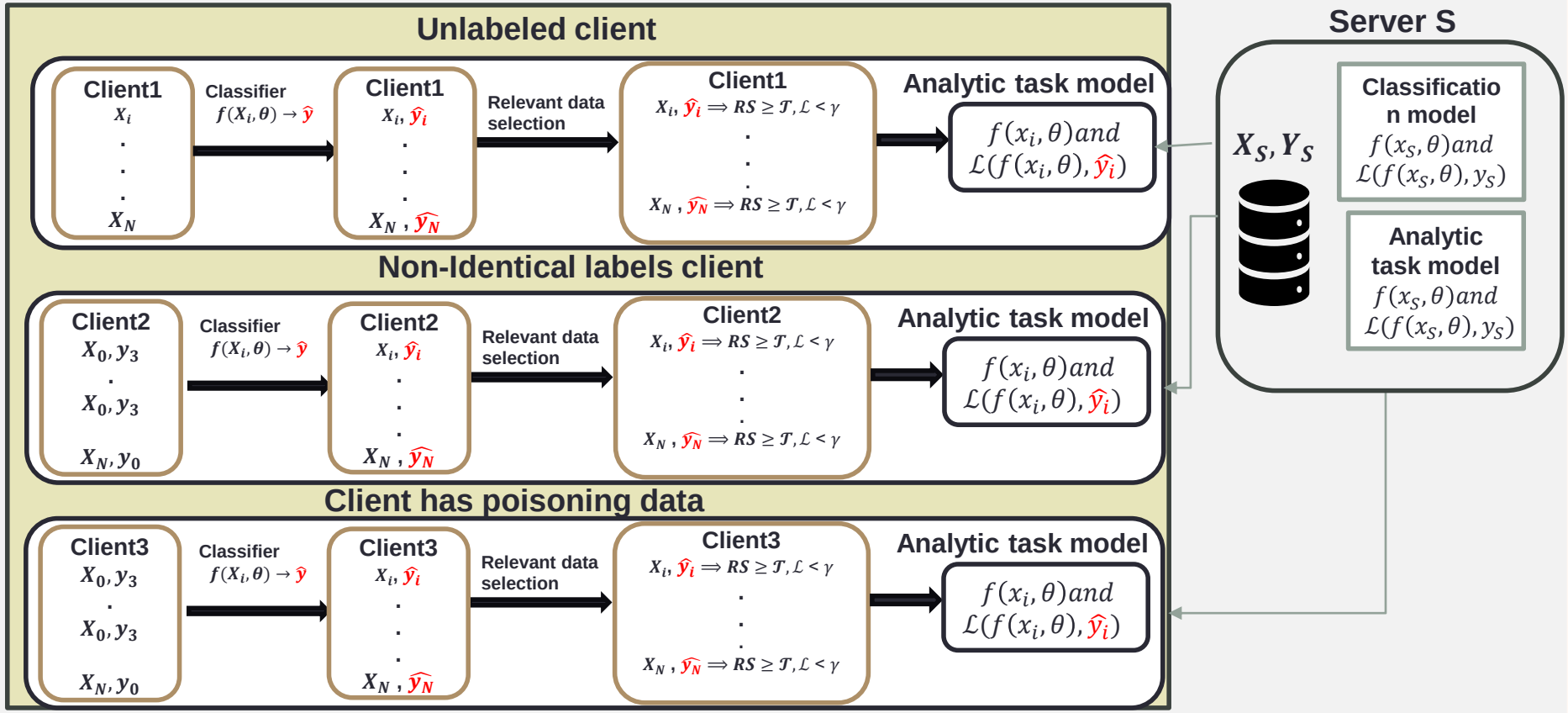
Leveraging the pseudo label $\hat{y}$ varies between unlabeled data, non- identically labeled data, and attacked data.

Common method: adopt a confidence threshold $\tau \in [0.5, 1]$ for the pseudo label probability:

$$\hat{y}_i = \begin{cases} 1 & g_{\theta_S}(x_i)c \in \tau \\ 0 & \text{otherwise} \end{cases}$$

- **Unlabeled data:** MP-SSFL considers the label if it satisfies τ condition
- **Non-identically labeled data**. MP-SSFL considers the sample's loss $\mathcal{L}\left(g\theta_S(x_i)\right)$, with γ tolerance threshold such that a label is accepted if $\mathcal{L} < \gamma$

MP-SSFL System Overview



Experiments

MP-SSFL is evaluated on datasets MNIST and Fashion-MNIST compared against:

- **Baseline** [1]: Relevant label selection relies on a probability determined by the Relevance Prediction Function Score for *each* sample.

- **Baseline** [2]: Label relevance selection is based on a model built only on server's data.

[1] T. Tuor, S. Wang, B. J. Ko, C. Liu, and K. K. Leung, "Overcoming noisy and irrelevant data in federated learning," 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 5020–5027.

[2] L. Nagalapatti, R. S. Mittal, and R. Narayanam, "Is your data relevant?: Dynamic selection of relevant data for federated learning," AAAI Conference on Artificial Intelligence, vol. 36, no. 7, 2022,pp. 7859–7867.

Experiments

MNIST									
Model	**Baseline1**			**Baseline2**			**MP-SSFL**		
Client	**10**	**50**	**100**	**10**	**50**	**100**	**10**	**50**	**100**
200 Sample	74.69	74.695	74.695	83.35	84.97	84.35	89.73	88.45	87.98
500 Sample	73.94	74.9	74.35	84.52	84.42	84.4	90.35	89.29	88.78
1000 Sample	74.39	74.369	74.56	84.63	84.24	84.36	91.56	90.87	91.24
F-MNIST									
200 Sample	68.75	68.05	67.47	76.7	77.79	74.78	84.11	85.05	84.93
500 Sample	67.45	68.09	67.9	74.05	74.76	74.33	85.37	85.71	86.27
1000 Sample	68.65	67.42	67.63	74.9	74.0	74.5	86.55	86.64	85.73


University of Glasgow
School of Computing Science
Knowledge & Data Engineering Systems
Thank you!