Federated learning as a tool for open models of machine learning in eGovernment

(TODO) Guberović, Emanuel; Čavrak, Igor; Bosnić, Ivana; Alexopoulos, Charalampos

Federated learning as a tool for open models of machine learning in eGovernment

Emanuel Guberović[1]; Igor Čavrak[2]; Ivana Bosnić[1]; Charalampos Alexopoulos[2]
[1] Sveučilište u Zagrebu, Faculty of electrical engineering and computing, Zagreb, Hrvatska
[2] University of Aegean, Greece

Abstract:
Federated learning (FL) emerged as a new data-parallel machine learning (ML) technique, contributing missing links needed in the field of artificial intelligence to comply with restrictions concerning data privacy regulations. Besides enabling ML to dodge data privacy obstacles, it creates new opportunities by facilitating global knowledge discovery through training models using distributed datasets from different data providers and with different ownership and access rights. Such an approach advocates the creation of open models – an extension of the open data concept – where data required for open model construction can be open, closed, and a combination of both. FL open models (FLOMs) align with the usage of new disruptive technologies for achieving 'knowledge of the crowd' in supporting data-driven and evidence-based decision and policy-making, recognized as a third-generation eGovernance methodology. This article proposes a simple FL framework, with a step-by-step guide on implementing a FLOM accompanied by two examples that fall within the eGovernance domain.
We specify the FLOM framework as a blueprint for using FL in realization of open models with the following specification items: client data and requirements, an aggregation server, an Application Programming Interface (API) on the aggregation server, and a runnable ML model. A high-level description of the required individual data and computational capabilities of the client for participation within the learning process includes required data attributes, their frequency, and quantity, as well as possible additional qualitative data metrics. An aggregation server is required to create an aggregate value from a set of model weight client updates, followed by successfully notifying and disseminating the new global model weights to the participating clients. The API interface on the aggregation server consists of endpoints for receiving client model weight updates and disseminating the new global weights. Notably, the ML model used at the core of the FLOM process needs to take the predefined input values from the client and provide the appropriate model weights for the API endpoint on the aggregating server. FLOM is based on the typical FL process that takes four distinct steps per one iteration: in the first step, clients send their individual model updates, followed in the second step by aggregation of those updates on the aggregation server. The third step requires returning the aggregated model weights to the clients, who use that data in the final step to update their local models.
We validated the potential of FLOM as a 3rd generation eGovernance tool using two different use cases; by comparing the quality of the data discovery with the confidential and private

data available to the FL process and using only the data available to the typical centralized ML. The first use case revolves around a horizontally partitioned environment, with a goal of agricultural commodity price prediction by combining data from the EUROSTAT price index and FAO product import/export dataset. This data is partitioned on a country level, with each one being a distinct data unit. Using FLOM in this example allows individual producers to gain better information about the cost-effectiveness of producing each commodity. This new knowledge can be discovered without the need for producers to exchange their production cost data, often confidential. The second use case relies on the constructed dataset from the anonymized private data created for a loan approval task containing credit record data and some client-specific private data. By vertically separating the dataset into credit balance data and private data, we compare the gains achieved using FL with the knowledge extracted from the complete dataset versus using only the credit balance data.

Our validation of FL and open model approach, based on the two use cases from the domain of eGovernance, revealed significant gains compared to using the data available only to centralized ML techniques. With the introduction of the FLOM framework, we aim to facilitate the creation of new tools, services, and usage scenarios from various domains that were previously not practically possible or hard to achieve. In particular, we aim at usage scenarios that would allow the creation of new knowledge, in the form of open models, that combine both open and closed datasets and allow various parties to participate in the creation and usage of such open models.

Keywords: Federated learning, machine learning, open data, open models, eGovernance