Conditioning on Local Statistics for Scalable Heterogeneous Federated Learning

Characteristic statistics such as means, covariances, and higher moments are calculated independently by each client using their own training dataset and appended to each data point. This approach enables the model to condition on the local data distribution in a federated learning setting, effectively handling distribution shifts between clients as shown in our numerical experiments. As no additional data is shared, the approach is privacy-preserving and scales well.

Rickard Brännvall
RISE Research Institutes of Sweden

Federated learning is a distributed machine learning approach where multiple clients collaboratively train a model without sharing their local data, which contributes to preserving privacy. **A challenge in federated learning is managing heterogeneous data distributions across clients**, which can hinder model convergence and performance due to the need for the global model to generalize well across diverse local datasets.

We propose to use local characteristic statistics, by which we mean some statistical properties calculated independently by each client using only their local training dataset. These statistics, such as means, covariances, and higher moments, are used to capture the characteristics of the local data distribution. They are not shared with other clients or a central node.

During training, the local statistics help the model learn how to condition on the local data distribution, and during inference, they guide the client's predictions. This allows for efficient handling of heterogeneous data across the federation and has favorable scaling compared to approaches like Clustered Federated Learning (CFL) that directly try to identify peer nodes that share distribution characteristics or Personalized Federated Learning (PFL) that finetunes the global model to each client's data. It also maintains privacy as no additional information is communicated.

	global	cluster	client	cond
linreg (rmse)	14.901	0.1	0.106	0.104
logreg (acc)	0.7	0.997	0.944	0.989
emnist (acc)	0.847	0.97	0.88	0.967

Table 1: Performance comparison of conditional models with reference models on three tasks.

Method

1. **Preparation**: Each client independently calculates local statistics μ using their own training data. While clients agree on the method, the resulting statistics are not shared.
2. **Training**: Each client inputs their local statistics μ alongside other training data. FedAvg or FedSGD can be used.
3. **Inference**: The local client uses its own static characteristics μ to tailor predictions to its specific data distribution during inferenc.

Experiments

- **Synthetic Tasks**: Three clusters with 100 clients each were used to generate feature vectors and regression coefficients. Three local conditioning models were evaluated on RMSE and accuracy for linear and logistic regression tasks.
- **EMNIST Task**: A three-layer CNN was trained on handwritten characters from the EMNIST dataset, using local characteristic stats μ calculated as the first principal component loading of the flattened image and label.

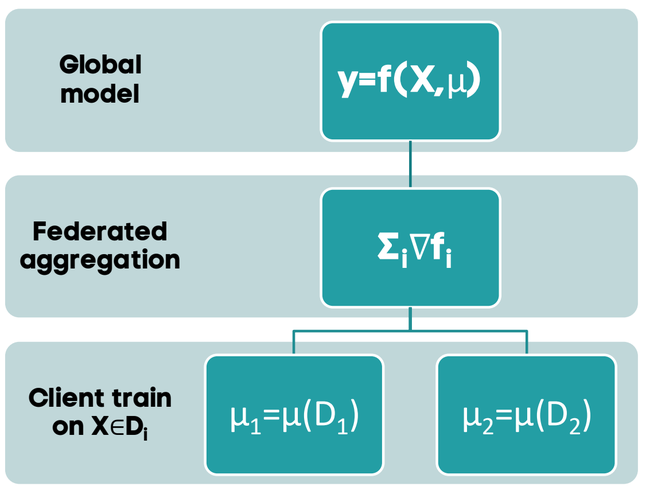

Figure 1: Federated learning set-up.

Results

On both tasks, the proposed conditioning models, outperforms both global and client-specific models and achieves model performance at par with the cluster-specific models that assume oracle knowledge of client peers. In the EMNIST character recognition task, the Conditional CNN achieves accuracy comparable to cluster-specific models, especially for challenging similar character sets like (z, Z, 2) and (i, I, 1). Global and client-wise models underperform due to lack of data and inability to handle heterogeneous distributions effectively.

Future work

- Examine more complex tasks and larger data sets.
- Compare performance to alternative methods that handle heterogenous data such as PFL and CFL.
- Investigate compression techniques for local statistics, e.g., latent embeddings for images.

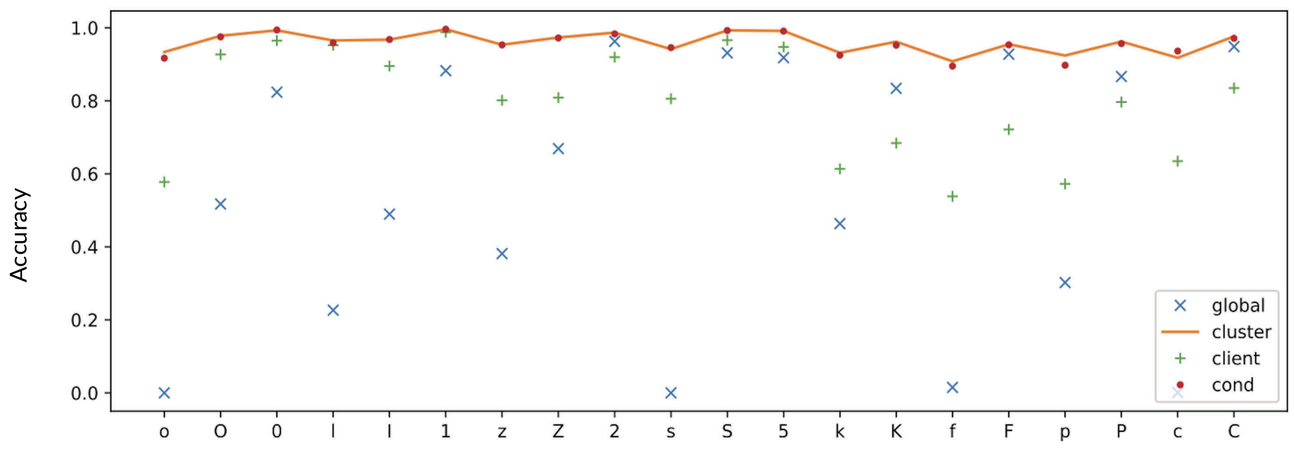

Figure 2: Conditional CNN performs better on similar EMNIST characters compared to global and client unconditional reference models. Its accuracy is at par with the cluster oracle model.

ABOUT RISE

RISE Research Institutes of Sweden is an independent, State-owned research institute with over 3000 employees, which offers unique expertise and over 130 testbeds and demonstration environments for future technologies, products and services.

RISE Research Institutes of Sweden
www.ri.se / info@ri.se

CONTACT
Rickard Brännvall
rickard.brannvall@ri.se