Optimisation de l'Apprentissage Fédéré en présence de données hétérogènes et déséquilibrées

Direction : Sonia Guehis, Inès Alaya, Sana Ben Hamida

Sujet de thèse au LAMSADE-Université Paris Dauphine PSL

1 Contexte et problématique

L'apprentissage fédéré (FL) [1], [2] permet d'entraîner des modèles de machine learning de manière décentralisée, sans partager directement les données brutes des participants. Ce paradigme est particulièrement intéressant pour les domaines où la confidentialité est primordiale (santé, finance, IoT, biodiversité marine, etc).

Cependant, les données utilisées dans l'apprentissage fédéré (FL) posent plusieurs défis majeurs, notamment leur nature non indépendante et non identiquement distribuée (Non-IID Data). Ces défis incluent l'hétérogénéité des données, le déséquilibre entre les volumes et les classes, les biais, ainsi que la disponibilité et la variabilité temporelle des données. Ces problèmes compromettent la convergence, la robustesse et la généralisation des modèles fédérés, d'où la nécessité d'optimiser les stratégies de traitement des données et d'agrégation des modèles.

2 Objectifs

Ce projet de thèse vise à améliorer l'apprentissage fédéré en agissant sur deux axes principaux.

Le premier axe concerne l'harmonisation des contributions locales. Pour cela, plusieurs approches sont envisagées. Tout d'abord, le développement de méthodes de ré-échantillonnage intelligent des données permettra de réduire le déséquilibre des ensembles locaux. Ensuite, l'utilisation de techniques d'augmentation de données contribuera à accroître la diversité et la représentativité de ces ensembles [3]. Enfin, des stratégies de pondération des échantillons seront mises en œuvre en se basant sur des mesures de qualité et de représentativité.

Le second axe porte sur l'optimisation des stratégies d'agrégation et de pondération des modèles. À cet effet, l'intégration de métaheuristiques, telles que les algorithmes évolutionnaires [4], permettra d'ajuster dynamiquement les contributions des clients. De plus, l'exploration de l'optimisation par colonies de fourmis [5] sera envisagée afin d'optimiser les stratégies d'agrégation en prenant

en compte la diversité et la répartition des données. Enfin, la mise en place de mécanismes adaptatifs tiendra compte de la qualité et de la diversité des données locales pour affiner encore davantage les performances de l'apprentissage fédéré.

3 Contact et candidature

Merci d'envoyer par e-mail un CV, une lettre de motivation, des relevés de notes, et éventuellement une lettre de recommandation à sana.mrabet@dauphine.psl.eu, sonia.guehis@lamsade.dauphine.fr et ines.alaya@parisnanterre.fr

References

[1] Betul Yurdem et al. "Federated learning: Overview, strategies, applications, tools and future directions". In: *Heliyon* 10.19 (2024), e38137. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2024.e38137. URL: https://www.sciencedirect.com/science/article/pii/S2405844024141680.

[2] Qiang Yang et al. "Federated Machine Learning: Concept and Applications". In: *CoRR* abs/1902.04885 (2019). arXiv: 1902.04885. URL: http://arxiv.org/abs/1902.04885.

[3] Yudith Cardinale, Sonia Guehis, and Marta Rukoz. "Classifying Big Data Analytic Approaches: A Generic Architecture". In: *Software Technologies*. Ed. by Enrique Cabello et al. Cham: Springer International Publishing, 2018, pp. 268–295. ISBN: 978-3-319-93641-3.

[4] Alain Petrowski and Sana Ben Hamida. *Evolutionary algorithms*. en. Wiley, 2017.

[5] Imen Ben Mansour, Ines Alaya, and Moncef Tagina. "A gradual weight-based ant colony approach for solving the multiobjective multidimensional knapsack problem". In: *Evolutionary Intelligence* 12.2 (June 2019), pp. 253–272. ISSN: 1864-5917. DOI: 10.1007/s12065-019-00222-9. URL: https://doi.org/10.1007/s12065-019-00222-9.