Appel A Projets Recherche 2023

Fédération de Recherche Agorantic
«Culture, Patrimoines, Sociétés Numériques »

Attention :

- Ne pas dépasser 5 pages
- Transmettre le fichier au format PDF intitulé : ACRONYME-AAP-blc-Agorantic-2023
- Envoyer le fichier à agorantic@univ-avignon.fr **avant le 24 mars 2023.**

Titre	apprenTissage fédéRé poUr deS données héTérogènes et sensibles (TRUST)
Acronyme	TRUST
Nom du/des porteur(s)	Anna Melnykova (LMA), Rachid Elazouzi (LIA), Pierre Henri (LBNC)
Laboratoires associés	LMA, LIA, LBNC
Budget demandé	8000 Euros
Résumé **Max. 1 000 caractères espaces compris**	L'objectif du projet est de traiter une problématique majeure liée aux données utilisées par l'intelligence artificielle. La projet propose d'utiliser l'apprentissage fédéré comme une solution afin de respecter des directives modernes sur la confidentialité des données (RGPD) de l'UE. Par contre l'apprentissage fédéré (FL) est toujours confronté au défis de l'hétérogénéité statistique qui peut produire un modèle global moins performant. Le projet étudiera quantitativement l'impact de l'hétérogénéité statistique sur FL et proposera des solutions efficaces pour améliorer ses performances tout en maintenant la confidentialité des données. Un cas d'étude sera réalisé sur la BREF et d'autres bases de données complémentaires portées par des membres de la FR Agorantic.

1. Contexte, positionnement, objectif(s)

Le projet TRUST est le résultat d'un travail initié pour préparer un sujet de thèse dans le cadre de l'appel à contrat doctoral Agorantic. En travaillant sur ce sujet, nous avons identifié différentes pistes d'étude sur l'apprentissage fédéré et les données hétérogènes.

Ces dernières années, avec la floraison d'applications et de services basés sur l'apprentissage machine (ML), la garantie de la confidentialité et de la sécurité des données est devenue une obligation critique. Les fournisseurs de services basés sur l'apprentissage machine sont non seulement confrontés aux difficultés de la collecte et de la gestion des données provenant de sources hétérogènes, mais aussi à celles de la mise en conformité avec des réglementations rigoureuses en matière de protection des données, telles que le règlement général sur la protection des données (RGPD) de l'UE. L'apprentissage automatique basé sur des approches centralisées est toujours associé à des risques de longue date en matière de protection de la vie privée liés à la fuite, à l'utilisation abusive et à l'abus de données personnelles. En parallèle, la fédération Agorantic s'est lancée à travers des appels à projet "Open Data" sur le partage des données ouvertes et pour permettre leur réutilisation par les chercheurs de la FR, d'autres chercheurs et les professionnels concernés. Dans un contexte juridique qui alimente le débat juridique émergent entre les contraintes de transparence des données publiques et le respect des

données personnelles, l'open data peut poser à court terme des problèmes liés aux données personnelles.

Dans le domaine de l'intelligence artificielle, Il existe un intérêt croissant pour un nouveau paradigme de ML distribué appelé Federated Learning (FL) **[La17]**, dans lequel les clients calculent leurs gradients locaux et les communiquent à un serveur central. Ce serveur centralisé orchestre ensuite des cycles d'apprentissage sur de grands volumes de données créés et stockés localement dans un grand nombre de clients. Cette procédure d'apprentissage se répète jusqu'à ce qu'un certain critère soit atteint. Cela permet aux clients participants de protéger leurs données et de résoudre les problèmes de sécurité et de confidentialité des données imposés par la loi. En effet, comme l'ont déjà suggéré plusieurs autorités de protection des données comme GDPR, FL a le potentiel de faciliter la conformité avec le principe de confidentialité des données. Cela est particulièrement important dans les applications de santé ou politique où les données regorgent d'informations personnelles et hautement sensibles, et où les méthodes d'analyse des données doivent probablement être conformes aux directives réglementaires. Les systèmes FL devraient connaître une croissance exponentielle, chaque système contenant lui-même un grand nombre de petits appareils dans différentes régions géographiques. En outre, les GPU puissants sont de plus en plus accessibles, ce qui permet de déployer des modèles plus importants, ce qui accélère le déploiement du FL. Cette demande croissante pour la technologie FL va ouvrir de nouveaux défis en plus de ceux qui apparaissent dans la ML traditionnelle.

2. Problématique, questionnement scientifique, axe(s) de la FR concerné(s)

La demande croissante pour la technologie FL ouvre de nouveaux défis en plus de ceux qui apparaissent dans la ML traditionnelle. Il s'agit notamment de l'allongement de la durée de la période d'apprentissage en raison de la corrélation entre les bases de données des clients. La plupart des formulations existantes de l'apprentissage fédéré le traitent comme un problème d'optimisation où la fonction de perte globale est optimisée sur plusieurs tours, chaque tour consistant en une estimation ponctuelle d'une fonction de perte définie sur les données locales du client, suivie d'une agrégation des modèles du client, puis d'une agrégation des modèles du client sur un serveur central. Par contre l'apprentissage fédéré est toujours confronté au défis de l'hétérogénéité statistique. Plus précisément, l'hétérogénéité statistique résulte des données non IID générées par différents clients, qui possèdent diverses caractéristiques ou distributions de probabilité d'étiquettes. Il est prouvé qu'elle a un impact négatif sur la convergence et la précision du modèle par rapport aux données homogènes (independent and identically distributed). Par exemple, si les données des clients sont à longue queue et hétérogènes à la fois, le problème commun devient compliqué et difficile car chaque client peut détenir différentes classes de queue. Cependant, le déséquilibre des données hétérogènes entre les clients, la limitation de la puissance de la batterie, des capacités de calcul et de communication, peuvent produire un modèle global biaisé ou une solution injuste entre les clients. Pour remédier à cette situation, nous identifions les problèmes liés aux pratiques actuelles et proposons des remèdes concrets en définissant une nouvelle notion de cadre d'hétérogénéité des données dans FL qui facilite davantage les évaluations standardisées et la comparaison des méthodes. Car les données viennent d'une ou plusieurs familles des distributions probabilistes, en définissant la hétérogénéité comme la différence entre ces distributions, nous pouvons se servir des outils probabilistes puissants pour mesurer l'impact sur la performance des modèles FL. La difficulté principale c'est que nous n'avons

accès qu'à certains descriptifs des données en question, mais jamais aux échantillons complets, c'est qui complique l'application des métriques standards (ex. Kullback-Leibler, ou la distance de Wasserstein).

Les approches développées dans la littérature manquent d'une élaboration approfondie sur le type d'hétérogénéité des données et sur la façon dont l'hétérogénéité des données affecte la précision des performances des clients participants. Dans ce projet, nous prévoyons d'incorporer des analyses statistiques sur les données des utilisateurs afin d'identifier les stratégies de sélection des clients qui peuvent bénéficier aux participants à l'apprentissage fédéré et réduire le temps d'apprentissage. Cependant, il existe un compromis entre la performance des systèmes fédérés obtenue grâce à ces stratégies et la confidentialité des données.

Le projet TRUST s'inscrit à l'intersection de deux axes de la Fédération de recherche : **AXE 1** "Méthodologies et interdisciplinarité" et **AXE 5** "Structuration et exploitation de corpus". L'apprentissage centralisé et la science des données sont des outils essentiels pour la science grâce à l'accessibilité croissante de la collecte, du stockage et du traitement de grandes quantités de données. Mais il y a des risques juridiques si les données ne sont pas correctement gérées. L'apprentissage fédéré peut apporter une réponse efficace aux problèmes d'utilisation des données et ouvre des perspectives intéressantes sur la possibilité de partager des données tout en préservant la confidentialité. Par contre les améliorations apportées à l'apprentissage fédéré doivent être soigneusement conçues si l'objectif est de fournir des garanties plus formelles telles que la confidentialité différentielle. C'est exactement l'objectif du projet qui projette d'étudier quantitativement l'impact de l'hétérogénéité statistique sur FL et proposer des solutions efficaces pour améliorer ses performances tout en maintenant la confidentialité des données.

3. Méthodologie

Les données sont traitées comme les observations d'une ou plusieurs familles des distributions probabilistes. En définissant l'hétérogénéité comme la différence entre ces distributions, nous pouvons se servir des outils probabilistes puissants pour mesurer l'impact sur la performance des modèles FL. La difficulté principale c'est que nous n'avons accès qu'à certains descriptifs des données en question, mais jamais aux échantillons complets, c'est qui complique l'application des métriques standards (ex. Kullback-Leibler, ou la distance de Wasserstein).

Dans le projet, nous allons étudier plusieurs pistes pour atteindre nos objectifs scientifiques:

1. Nous étudierons quantitativement l'impact de l'hétérogénéité statistique sur l'apprentissage fédéré,
2. nous définirons des métriques (ex. Kullback-Leibler) permettant d'évaluer l'hétérogénéité statistique des données tout en garantissant la confidentialité des données,
3. nous développerons une stratégie pour sélectionner les clients pendant l'apprentissage en utilisant la métrique qui mesure l'hétérogénéité statistique de chaque client, ce qui peut limiter efficacement la divergence des modèles pendant l'apprentissage fédéré. Par contre, nous aurons toujours le choix entre l'efficacité des méthodes proposées et le risque de confidentialité des données. Une des approches qui va être explorée dans cette thèse est le réseau neuronal bayésien, introduit dans FL pour résoudre le problème de

l'overfitting du modèle en représentant tous les paramètres du réseau dans le modèle global avec des distributions de probabilité[1].

Nous allons se servir des outils probabilistes pour mesurer l'hétérogénéité des données, et puis appliquer les méthodes de FL pour proposer une méthode d'apprentissage qui ne compromet pas la confidentialité des données.

Un autre aspect que nous souhaitons étudier dans ce projet est l'hétérogénéité des objectifs de chaque client participant à l'apprentissage fédératif. La plupart des algorithmes d'apprentissage fédératif visent à apprendre un modèle global qui atteint une performance globale optimale pour tous les clients. Le déséquilibre des données hétérogènes entre les clients peut produire un modèle global moins performant. De nombreux travaux antérieurs ont souligné que l'hétérogénéité des données statistiques dans FL a des effets néfastes et peut conduire à une mauvaise convergence[2][3] ce qui nécessite une personnalisation[4]. La majorité des travaux dans la littérature ne sait toujours pas si et dans quelles conditions les clients bénéficient de la collaboration dans un contexte d'hétérogénéité des données[5]. Cela signifie que chaque client est confronté à un choix difficile : doit-il participer au système FL ou rester avec son modèle local ? Les travaux existants partent de l'hypothèse que tous les clients participeront inconditionnellement au système FL sans connaître l'avantage de rejoindre un FL pendant la période de formation. Ici, nous étudions un nouveau contexte de FL en donnant à chaque client l'opportunité de mesurer le gain en termes de précision en client avec un sous-ensemble de client. Ainsi, l'un des principaux objectifs est de quantifier l'influence de chaque base de données en termes de paramètres du modèle et de proposer de nouveaux algorithmes d'estimation efficaces sous contraintes de confidentialité. L'une des approches concerne l'élaboration d'une nouvelle présentation d'une donnée locale basée sur des techniques de résumé de données. Ainsi, un mécanisme d'incitation est nécessaire pour construire un FL flexible en utilisant un jeu de formation de coalition, qui est l'une des approches fondamentales dans les systèmes multi-agents pour établir une collaboration entre des groupes ou des clients intéressés[6]. Ce projet se positionne dans le développement d'approches efficaces pour mesurer la contribution de chaque client. Le facteur décisif pour obtenir des algorithmes efficaces est que toutes les coalitions ne doivent pas être prises en compte dans le calcul de la précision.

4. Résultats attendus et caractère innovant de la recherche

L'apprentissage fédéré est un domaine de recherche actif et continu. Bien que des travaux récents aient commencé à aborder les défis discutés dans ce projet, il reste encore un certain nombre de

[1] **[Chen21]** Chen, H.-Y. and Chao, W.-L. FedBE: Making Bayesian model ensemble applicable to federated learning. In Inter-national Conference on Learning Representations, 2021.
[2] **[Fallah20]** Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. Advances in Neural Information Processing Systems, 33:3557–3568, 2020.
[3] **[Dennis21]** Don Kurian Dennis, Tian Li, and Virginia Smith. Heterogeneity for the win: One-shot federated clustering. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139, pages 2611–2620. PMLR, 2021.
[4] **[Li21]** Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. CVPR, abs/2103.16257, 2021.
[5] H Kabbaj, M El-hanjri, A Kobbane, R El-Azouzi, "DistFL: An enhanced FL approach for Non Trusted Setting in Water Distribution Networks" 2023 IEEE International Conference on Communications (ICC)
[6] M. Touati, **R. El-Azouzi**, M. Coupechoux, E. Altman and J. Kelif, "A Controlled Matching Game for WLANs," in *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 3, pp. 707-720, March 2017

directions critiques à explorer. En effet, le projet vise à quantifier l'hétérogénéité statistique et son importance dans l'apprentissage fédéré. Le résultat attendu de ce projet est de fournir des solutions pour quantifier le degré d'hétérogénéité lié aux systèmes afin de l'exploiter pour améliorer la convergence des méthodes d'optimisation des réseaux fédérés. Cependant le projet aborde un problème intéressant pour l'apprentissage fédéré à l'aide d'approches fondées sur les statistiques et la théorie des jeux.

5. Dimension interdisciplinaire (champs disciplinaires associés) et cohérence par rapport à la thématique « Culture, Patrimoines, Sociétés Numériques »

Ce projet est par construction interdisciplinaire. Il nécessite une triple expertise en informatique, en mathématiques et en droit. L'informatique apporte toute l'expertise sur l'apprentissage fédéré. Les mathématiques fourniront une analyse théorique sur des métriques à explorer pour répondre aux défis liés à l'hétérogénéité des données. Le droit joue un rôle important sur l'aspect lié aux contraintes imposées par le RGPD de l'UE. Enfin, nous prévoyons de mettre en place un apprentissage fédéré utilisant plusieurs bases de données portées par Agorantic pour tester nos solutions sous les contraintes du RGPD.

6. Partenariats extérieurs envisagés

Nous prévoyons collaborer avec des chercheurs à l'université de Carnegie Mellon à Pittsburgh (Prof. Giulia Fanti et Prof. Virginia Smith) et India Institute of Technology, Bhilai (Prof. Arzad Alam Kherani). Les trois chercheurs travaillent activement sur le sujet et cette collaboration permet de renforcer notre contribution dans ce domaine.

Budget (€)*		
	Brève description	**Montant**
Missions	On prévoit de publier les résultats sur les conférences internationales, le budget doit couvrir les frais d'inscription et les dépenses liés au trajet et à l'hébergement.	1000 Euros
Chercheurs non-permanents	2 Stages de durée de 6 mois Séjour d'un doctorant travaillant sur le même sujet (3 mois)	7000 Euros 2500 Euros
Budget total		9500 Euros
Co financements le cas échéant	ANR	1500 Euros
Budget demandé à Agorantic		**8000 Euros**