

Titre de la thèse : Vers une meilleure évaluation des modèles de machine learning

Laboratoire : LIMOS, Clermont-Ferrand

Contexte :

Il existe de nombreux algorithmes de machine learning pour lesquels leurs avantages et inconvénients sont globalement connus. Néanmoins, il reste difficile de savoir quel algorithme utiliser pour un jeu de données spécifique, car les guides d'utilisation des algorithmes incluent seulement les caractéristiques les plus simples des jeux de données (nombre d'attributs, nombre de classes, etc.). Or les caractéristiques d'un jeu de données sont bien plus complexes et ont une influence non négligeable sur la performance des algorithmes. Cette thèse vise à mieux renseigner les caractéristiques pour lesquelles un algorithme de machine learning (non supervisée dans un premier temps, supervisée dans un second temps) aurait de bonnes/mauvaise performances.

Sujet :

Les objectifs de la thèse sont les suivants :

- Proposer des formules et définir des métriques pour caractériser des jeux de données. Les avantages à mesurer ces caractéristiques sont nombreux. Par exemple, elles peuvent aider un chercheur à mieux présenter les avantages et les inconvénients de nouveaux algorithmes de classification.
- Associer un type de jeu de données défini par ses caractéristiques à un ensemble d'algorithmes de classification. La finalité serait de créer des guides d'utilisation qui permettent, dans le cadre d'un problème appliqué avec un jeu de données spécifique, de choisir le ou les algorithmes les plus intéressants ;
- Proposer des méthodes pour générer de nouveaux jeux de données selon les propriétés qu'un utilisateur cherche à vérifier. Comme certains jeux de données peuvent ne pas être suffisants pour analyser certaines caractéristiques, il est également nécessaire de créer des algorithmes qui génèrent des données synthétiques afin de combler le manque de jeux de données avec certaines caractéristiques;
- Etudier la robustesse des modèles à un pourcentage d'erreur ajouté aux données (sous forme d'outliers ou d'étiquettes bruitées)

Financement : Ce sujet de thèse est éligible aux financements doctoraux du Ministère de l'Enseignement supérieur et de la Recherche. Les candidat·e·s doivent impérativement figurer dans le tiers supérieur de leur promotion pour prétendre à ce type de financement.

Formation et compétence requises : Etudiant diplômé en BAC+5 avec une spécialisation en informatique, des connaissances en science des données, et un intérêt pour les mathématiques.

Candidature : La candidature doit contenir un CV, ainsi que le relevé de notes de BAC+4 et les notes déjà obtenues en BAC+5. La date limite pour candidater est le **06/04/2026**

Contacts : Violaine Antoine violaine.antoine@limos.fr, Sébastien Salva sebastien.salva@limos.fr