Projet de thèse : Exploration de méthodes d'assemblage de modèles pour la prédiction en spectroscopie proche infrarouge

Mots-clés

Spectroscopie proche infrarouge, apprentissage automatique, empilement de modèles, intelligence artificielle, calibration, stacking

Laboratoire d'accueil

UMR AGAP Institut,Avenue Agropolis - 34398 Montpellier Cedex 5 (https://umr-agap.cirad.fr/)

- **Directeur de thèse :** Fabien Michel (HDR), LIRMM, Équipe SMILE

- **Encadrants**

Grégory Beurier, CIRAD, AGAP Institut
Lauriane Rouan, CIRAD, AGAP Institut
Denis Cornet, CIRAD, AGAP Institut

Contexte et problématique de l'étude

La spectroscopie proche infrarouge (NIRS) est une technique d'analyse rapide, non destructive et à faible coût, très largement utilisée dans de nombreux domaines tels que la santé, la chimie, l'agro-alimentaire et notamment l'agronomie. Elle permet de déterminer la composition chimique et les propriétés fonctionnelles d'échantillons de produits tels que les grains, fourrages, aliments, et tissus. Les données spectrales générées par NIRS sont riches en informations mais nécessitent des traitements statistiques avancés pour des prédictions précises. Des méthodes comme la régression PLS ont été historiquement utilisées, mais les avancées en apprentissage machine (réseaux de neurones, SVM, random forest, etc.) et l'accès à d'importantes bases de données NIRS ont permis l'adoption croissante de ces méthodes d'intelligence artificielle, qui démontrent souvent de meilleures performances prédictives.

La démocratisation des spectromètres et l'augmentation croissante d'utilisateurs non spécialistes, au Nord comme au Sud, renforce la nécessité de développer une approche générique et performantes de la calibration de modèles NIRS. Le stacking, méthode qui combine les prédictions de multiples modèles, a démontré son potentiel pour exploiter les forces complémentaires de différents algorithmes et améliorer les performances de prédiction. Cependant, les stratégies de stacking restent sous-explorées pour l'analyse des données NIRS. Dans ce contexte, le package Python Pinard (a Pipeline for Nirs Analysis ReloadeD, https://pypi.org/project/pinard/) développé par l'équipe encadrante fournit une base idéale pour l'implémentation et le test des approches de prédiction à base de stacking.

Objectifs du projet de thèse

L'objectif principal de cette thèse est de développer et optimiser des stratégies de stacking adaptées à la prédiction à partir de spectres NIRS en s'appuyant sur le package Pinard. Pinard fournit déjà des outils pour le traitement et l'analyse des données NIRS, y compris des modèles prédictifs individuels, mais ne propose pas actuellement de méthodologies d'assemblage de modèles. Cette recherche vise à combler cette lacune en intégrant des techniques avancées de stacking, permettant une amélioration significative des performances prédictives.

En particulier, le travail de thèse s'articulera autour des axes suivants (qui peuvent évoluer en cours de doctorat et qui sont d'importances variés) :

• **Axe 1 :** étudier et concevoir des méthodes de standardisation des données pour nourrir les différentes classes de modèles de la stack ; en particulier en ce qui concerne les contraintes des modèles de machine learning ou des sources différentes. Ce travail inclura également une prise en main et une analyse poussée des jeux de données à disposition.

• **Axe 2 :** sélectionner, intégrer et hyperparamétrer des modèles de prédictions (existants ou nouveaux) au sein d'une stack « traditionnelle » et étudier l'impact de chacun sur la précision globale en fonction des jeux de données et des méthodes d'assemblage (sélection aléatoire, sélection basée sur la performance, sur la diversité des algorithmes, sur la dissemblance des prédictions, etc.).

• **Axe 3 :** Concevoir et explorer des stratégies efficaces afin d'améliorer les stratégies de stacking de modèles en termes de précision, d'efficacité et de sobriété :

- Heuristiques issues de l'intelligence artificielle distribuée (systèmes multi-agents) ou de l'optimisation (méthodes évolutionnistes),
- Calcul temps réel de la contribution et/ou de l'explicabilité des modèles,
- Organisation et sélection dynamique des prétraitements de données,
- Hyperparamétrisation partielle temps réel,
- Etc.

L'axe 3 est au cœur de la problématique de cette thèse et devrait légitimement représenter une grande partie du travail du doctorant.

• **Axe 4 :** Travailler sur la diffusion des résultats obtenus que ce soit en facilitant la réutilisation de la stack ou l'accès aux outils et méthodes :

- Transfert de modèles à de nouveaux analytes / jeux de données / machines,
- Etude de l'explicabilité sous-jacente des modèles de la stack et identification des composants du signal,
- Intégration des développements dans le package Pinard.

Ce travail fournira des approches innovantes et performantes pour exploiter la richesse des données NIRS. Ainsi, il permettra d'améliorer la précision et la robustesse des analyses NIRS pour des problématiques telles que l'identification rapide de variétés adaptées aux défis climatiques, la détection et quantification de contaminants biotiques et abiotiques dans les récoltes, l'optimisation de la qualité et la valeur nutritive des aliments transformés, etc.

contribuant de fait à des thématiques chères au CIRAD telles que la sécurité alimentaire, la gestion durable des ressources et l'amélioration de la santé dans les pays du Sud.

Matériels et Méthodes

Ce projet s'appuiera sur une base de données diversifiée déjà construite, comprenant 30 jeux de données issus de domaines variés, allant de l'analyse de l'amidon dans les tiges de sorgho à la teneur en substances actives dans les médicaments, en passant par le taux de gras dans la viande ou l'indice d'octane de fuels industriels. Cette richesse de données offre une opportunité unique de tester et d'appliquer les nouvelles méthodologies dans un large éventail de contextes.

Nous insistons également sur le fait que dans le cadre du travail sur ces données, une attention particulière sera portée à l'efficience énergétique des processus de calcul, avec l'objectif de minimiser l'impact énergétique comme conséquence bénéfique de l'optimisation des algorithmes. En effet, identifier efficacement les combinaisons de modèles conduisant aux meilleures prédictions, contrairement aux méthodes traditionnelles d'exploration exhaustive, permet de minimiser les ressources requises (temps, calcul, énergie). Cet aspect s'inscrit dans une démarche de développement durable et de responsabilité environnementale.

Résultats attendus

• **Axe 1 :** Etude et développement de méthodes efficaces de prétraitements des données, intégrées à Pinard, permettant d'entrainer toutes les classes d'algorithmes utilisées en stacking sur tous les formats de données NIRS disponibles.

• **Axe 2 :** Utilisation efficace des stratégies d'assemblage de modèles et amélioration significative des performances prédictives en comparaisons des résultats avec modèles uniques sur les données du jeu de référence. Des modèles de stack préconfigurées sont intégrées dans Pinard pour différents cas d'usage.

• **Axe 3 :** Exploration et conception de méthodes d'amélioration de l'efficacité du stacking (gain de temps d'apprentissage, diminution du nombre de modèles, amélioration des résultats, etc.). Les résultats liés à cet axe dépendront du succès des différentes stratégies qui seront développées et testées. Dans tous les cas, ces résultats seront intégrés à la librairie Pinard.

• **Axe 4 :** Hormis les publications attendues, les leviers de diffusion d'un tel travail concernent d'une part l'utilisabilité (en particulier au Sud) des méthodes développées et leur transférabilité. De ce point de vue les méthodes devront être utilisable facilement avec Pinard mais également pouvoir répondre à la diversité des problématiques existantes en NIRS (données labo, données champs, absorbance, réflectance, etc.), et d'autre part la compréhension qu'aura l'utilisateur des résultats, de ce point de vue, une attention particulière sera portée sur l'interprétabilité des résultats obtenus par le stacking.

Conditions scientifiques et matérielles

Ce projet de thèse est financé via le co-financement, AAP Emergence Région (50%) et un cofinancement du CIRAD (50%). Les frais de fonctionnement de la thèse et liés à la création des jeux de données sont couverts par des projets déjà en cours. Les calculs nécessaires à la réalisation de ces travaux seront réalisés sur des calculateurs locaux (Geforce(s) 4090) et les clusters Jean Zay et adastra.

Ouverture internationale

La participation à plusieurs conférences internationales permettra au doctorant de développer son réseau international. Des collaborations avec des partenaires du Sud impliqués dans l'acquisition des données pour les espèces tropicales seront également envisagées.

Collaborations envisagées

Au niveau de l'UMR AGAP Institut, des collaborations seront mises en place avec les équipes Phenomen et GSP, qui se mobilisent actuellement dans l'utilisation des approches de deep learning dans la prédiction génomique. A l'international, de nombreux partenariats sont envisageables afin d'adapter les développements de Pinard aux besoins des utilisateurs du Nord et du Sud (e.g. CNRS, INRAe, Université de Potsdam, Boyce Thompson Institute, IITA, CNRA, NRCRI).

Objectifs de valorisation

Les résultats seront valorisés sous forme de publications dans des revues à facteur d'impact et de présentations dans des colloques internationaux. Le travail sur l'optimisation d'une méthodologie récente permettra de valoriser relativement facilement les résultats de recherche, très attendus dans ce domaine. Un papier sur la base de données utilisée et son analyse est aussi envisageable ainsi que pour tout autre avancée technique ou méthodologique qui sera réalisée dans cette thèse.

Profil recherché

• **Diplômes requis :** Master en informatique, bioinformatique, mathématiques appliquées, statistiques, ou sciences agronomiques avec une spécialité data science.

• **Compétences requises :**
- Développement en Python
- Data science et/ou statistiques
- Anglais (lu, écrit, parlé)
- Connaissances en R (optionnel)
- Traitement du signal (optionnel)
- Appétence pour la pluridisciplinarité

Formation continue du doctorant

Les formations nécessaires seront choisies parmi les catalogues offerts par l'Université de Montpellier (École Doctorale I2S), AGAP Institut, le Cirad et l'INRAE selon les besoins spécifiques du projet.

Suivi d'avancement du projet de thèse

Un comité de suivi de thèse sera organisé chaque année pour assurer le bon déroulement de la thèse et réorienter certaines parties si nécessaire. Ce comité sera composé de membres dont les domaines d'expertises seront liés au sujet traité (Deep Learning, traitement du signal, algorithmique et calcul, Chimiométrie, etc.).

Pour assurer un suivi rigoureux du projet de thèse, nous organiserons une réunion mensuelle avec le directeur de thèse et les encadrants pour discuter de l'avancement global, des résultats obtenus, des obstacles rencontrés et des objectifs à venir. En complément, des réunions hebdomadaires avec les encadrants permettront de suivre de près les travaux en cours, d'apporter un soutien technique et d'ajuster les plans de travail. Un dépôt GitHub partagé sera utilisé pour centraliser et suivre les travaux, facilitant ainsi la collaboration, la traçabilité des modifications et le partage des progrès réalisés de manière efficace et sécurisée.

Informations administratives

- **Année universitaire de 1ère inscription en doctorat :** 2024
- **Date de début de la thèse :** 01/10/2024
- **Date limite de candidature :** 23h59 21/07/2024
- Pièces à fournir : CV, lettre de motivation
- Renvoyer l'ensemble des pièces à denis.cornet@cirad.fr

Contacts

- **Grégory Beurier** – beurier@cirad.fr
- **Lauriane Rouan –** lauriane.rouan@cirad.fr
- **Denis Cornet –** denis.cornet@cirad.fr

CIRAD Agricultural Research for Development
AGAP Mixed Research Unit "Genetic Improvement and Adaptation of Mediterranean and Tropical Plants".
Avenue Agropolis - 34398 Montpellier Cedex 5
France

- **Fabien Michel** – fmichel@lirmm.fr

LIRMM -Université de Montpellier - CNRS, Équipe SMILE
Montpellier, France
http://www.lirmm.fr/~fmichel

Publications pertinentes de l'équipe encadrante

1. Vasseur F. et al. (2022). A Perspective on Plant Phenomics: Coupling Deep Learning and Near-Infrared Spectroscopy. https://doi.org/10.3389/fpls.2022.836488
2. Houngbo M. E. et al. (2023). Convolutional neural network allows amylose content prediction in yam (Dioscorea alata L.) flour using near infrared spectroscopy. https://doi.org/10.1002/jsfa.12825
3. Alamu E. O. et al. (2020). Near-infrared spectroscopy applications for high-throughput phenotyping for cassava and yam: A review. https://doi.org/10.1111/ijfs.14773
4. Sambakhé D. et al. (2019). Conditional optimization of a noisy function using a kriging metamodel. https://doi.org/10.1007/s10898-018-0716-0
5. Bonnici I. et al. (2022). Input addition and deletion in reinforcement: towards protean learning. https://doi.org/10.1007/s10458-021-09534-6