

Sujet de thèse :

Approches exploratoire et prédictive pour l'identification des réponses des plantes aux vagues de chaleur récurrentes

Lydia Boudjeloud-Assala, Miguel Couceiro
LORIA (Orpailleur)
{lydia.boudjeloud,miguel.couceiro}@loria.fr

Contexte

Depuis quelques années maintenant, nous observons de plus en plus des températures records pouvant se produire localement, qui affecte gravement la production agricole en quantité et en qualité. Non seulement les périodes de hautes températures sont brutales mais leur fréquence tout au long du cycle de vie d'une plante augmente également, accentuant la notion de "stress thermiques récurrents". Cela engage les agronomes à trouver des moyens d'augmenter la valeur génétique des cultures pour gagner en tolérance à la chaleur. En amont de la sélection végétale proprement dite, il est nécessaire de mieux comprendre les réponses des plantes aux stress thermiques isolés et récurrents qui se produisent au cours des étapes cruciales du développement des cultures.

Chaque espèce de culture a ses propres maxima de température au-delà desquels les processus de développement et physiologiques des plantes sont inhibés par des températures plus élevées. Le terme «réponse au stress thermique» est utilisé pour exprimer les réponses aux chocs thermiques soudains, aux longues vagues de chaleur et, plus généralement, aux expériences de réchauffement qui varient en durée et en intensité. Le stress thermique affecte négativement la croissance et le développement et, par conséquent, la qualité et la quantité.

Le sujet de thèse rentre dans le cadre du projet ANR RICCOCHETS, dont l'objectif est d'identifier les réponses des plantes aux vagues de chaleur récurrentes, en prenant en compte les différents paramètres étudiés sur les plantes tels que le phénotype multi-échelle des plantes, les paramètres écophysiological, et autres paramètres.

L'objet de l'étude concerne les différentes séquences du stress thermique (isolées et récurrentes) qui affectent la production céréalière (rendement et qualité) en étudiant des ensembles de données de terrain existants et ceux issues des expériences dans des conditions contrôlées. Il s'agit principalement de la croissance de trois graines, de colza, de blé et de sorgho, trois cultures importantes qui se différencient sur plusieurs aspects : type (oléagineux et céréales), les saisons de culture (par exemple, floraison printanière ou estivale pour le sorgho), la croissance (qui peut être déterminée pour le blé et le sorgho, et de croissance indéterminée pour le colza),...

L'objectif du projet RICOCHETS est de mieux comprendre et de prédire les effets des vagues de chaleur ponctuelles et récurrentes sur la production céréalière en tenant compte à la fois de la quantité et de la qualité.

Les données à disposition ont été obtenues sur trois cultures différentes (sorgho, colza et blé), en identifiant des séquences de stress thermique (intensité, durée et position pendant les phases de reproduction) qui ont un impact sur la production végétale affectant soit le nombre de grains, la taille des grains ou les deux, avec ou sans affecter la composition biochimique des grains. En comparant les effets d'un stress thermique isolé avec ceux de séquences récurrentes de stress thermique.

Des études récentes ont mis en évidence que la réponse finale de la plante à une séquence d'événements stressants récurrents ne correspond pas à la somme des réponses individuelles à chaque événement.

Il s'agit donc dans un premier temps d'explorer les données et d'utiliser des approches non supervisées (clustering, bi-clustering [1, 2]) afin d'identifier l'impact du stress thermique (la chaleur) sur les paramètres de croissance et de production de la plante. Dans un second temps de combiner les approches sélectionnées, avec les approches de découverte de relations causales.

Une relation causale décrit la relation entre deux variables, où une variable nommée *traitement* a un effet sur une variable nommée *résultat*. Les relations causales sont intéressantes dans de nombreux domaines, comme par exemple en santé pour déterminer si un médicament traite ou non une maladie, ou en politiques publiques pour comprendre si une nouvelle loi a eu un impact attendu ou non. Lorsque des données observationnelles sont disponibles, comme dans notre cas ici avec les trois grains, l'utilisation du cadre d'étude et des résultats des bi-clusters potentiels permettra d'identifier l'impact d'un paramètre (ou plusieurs) sur la résistance au stress thermique d'une plante.

Il existe différents cadres pour étudier le problème de la découverte de relations causales à partir de données d'observation tabulaires, tels que le modèle causal structurel [3] ou le cadre des résultats potentiels [4]. d'autres approches ont également été explorés dans le cadre de la gestion du réseau électrique qui pourraient être utilisés dans le cadre de ce stage [5, 6].

Dans un second temps, une fois le comportement expliqué, il est demandé de proposer un modèle de prédiction afin de prévoir le comportement des plantes face aux vagues de chaleurs pour mieux appréhender ses phases. les données (entrées et sorties) pour cette partie seront issues de l'étape précédente. L'aspect temporel des données est également à prendre en compte.

Travail demandé

1. Etude bibliographique sur les approches de clustering, bi-Clustering, relations causales, les données temporelles et les données biologiques des plantes.
2. Une approche exploratoire de découverte des connaissances pour identifier les séquences de stress thermique isolées et récurrentes et leurs impacts sur la production céréalière est à développer. Des algorithmes automatiques de clustering ou de bi-clustering des environnements en fonction des séquences de stress thermiques tout au long du cycle de culture seront développés pour

identifier les séquences de stress thermiques ponctuels et récurrents et analyser leurs effets respectifs sur la production végétale. L'aspect explicatif sera proposé avec des approches de relations causales.

3. Développement d'une approche prédictive des réponses au stress thermique, selon les descriptions des caractéristiques des plantes découvertes en phase deux.

Compétences demandés

- Master 2 en informatique, avec des compétences en machine learning et data science.
- Des compétences en biologie seraient un plus.

Information supplémentaires

- Thèse rémunérée (Projet ANR).
- Lieu de la thèse : LORIA site de Metz.
- Environnement de travail favorable à la recherche, une équipe de chercheurs, doctorants, stagiaires sur le site de CentraleSupélec.

Pour candidater

Pour candidater, veuillez nous envoyer votre CV détaillé, vos relevés de notes des 3 dernières années, une lettre de motivation, une lettre de recommandation du directeur du stage de Master aux adresses :

- Miguel Couceiro (Pr, LORIA Nancy) : miguel.couceiro@loria.fr
- Lydia Boudjeloud-Assala (MCF-HDR, LORIA Metz) : lydia.boudjeloud@loria.fr

Références

- [1] Lydia Boudjeloud-Assala and Alexandre Blansché. Iterative evolutionary subspace clustering. In Tingwen Huang, Zhigang Zeng, Chuandong Li, and Chi-Sing Leung, editors, *Neural Information Processing - 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part I*, volume 7663 of *Lecture Notes in Computer Science*, pages 424–431. Springer, 2012.
- [2] Ibrahim Louhi, Lydia Boudjeloud-Assala, and Thomas Tamisier. Subspace clustering et visualisation des flux de données. In Fabien Gandon and Gilles Bisson, editors, *17ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2017, 24-27 Janvier 2017, Grenoble, France*, volume E-33 of *RNTI*, pages 327–332. Éditions RNTI, 2017.
- [3] Ang Li and Judea Pearl. Bounds on causal effects and application to high dimensional data. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 5773–5780. AAAI Press, 2022.

- [4] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5) :688–701, 1974.
- [5] Laure Crochepierre, Lydia Boudjeloud-Assala, and Vincent Barbesant. Approche interactive d’extraction de variables interprétables et explicatives pour la gestion des contraintes du réseau électrique français. In Jérôme Azé and Vincent Lemaire, editors, *Extraction et Gestion des Connaissances, EGC 2021, 25-29 Janvier 2021, Montpellier, France*, volume E-37 of *RNTI*, pages 373–380. Éditions RNTI, 2021.
- [6] Antoine Marot, Antoine Rosin, Laure Crochepierre, Benjamin Donnot, Pierre Pinson, and Lydia Boudjeloud-Assala. Interpreting atypical conditions in systems with deep conditional autoencoders : The case of electrical consumption. In Ulf Brefeld, Élisabeth Fromont, Andreas Hotho, Arno J. Knobbe, Marloes H. Maathuis, and Céline Robardet, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part III*, volume 11908 of *Lecture Notes in Computer Science*, pages 638–654. Springer, 2019.