Classification automatisée des groupes fonctionnels du phytoplancton à partir de sorties de cytométrie en flux

Oral

Robin Fuchs[1] , Melilotus Thyssen[2] , Gérald Gregory[3] , Denys Pommeret[4] , Samuel Soubeyrand[5]

(1) Institut de Mathématiques de Marseille (I2M), Mediterranean Institute of Oceanography (MIO) - Equipe CYBELE robin.fuchs@mio.osupytheas.fr

(2) Mediterranean Institute of Oceanography (MIO) - Equipe CYBELE melilotus.thyssen@mio.osupytheas.fr

(3) Mediterranean Institute of Oceanography (MIO) - Equipe CYBELE gerald.gregori@mio.osupytheas.fr

(4) Institut de Mathématiques de Marseille (I2M) denys.pommeret@univ-amu.fr

(5) INRIA Avignon – Equipe BioSP samuel.soubeyrand@inria.fr

L'océanographie du fait de ses jeux de données massifs et en grande dimension, *i.e.* comportant un grand nombre d'observations et de variables, est un terrain privilégié pour l'analyse statistique.

C'est particulièrement le cas des données relatives au phytoplancton. Le phytoplancton, sous une désignation unique, comporte en réalité plusieurs milliers d'espèces. On peut néanmoins discerner plusieurs groupes d'espèces présentant d'importantes similarités (taille, temps de division etc...), appelés groupes fonctionnels.

Le développement de la cytométrie en flux a rendu possible l'acquisition automatisée de données sur les groupes fonctionnels du phytoplancton (un cytomètre pouvant traiter jusqu'à 10 000 cellules par seconde). A partir de chaque cellule de l'échantillon, le cytomètre est capable de générer un jeu de 5 courbes qui représente le profil optique de chaque cellule. Après retraitement, les océanographes déterminent ensuite manuellement les différents groupes fonctionnels présents dans les données. Ces opérations sont cependant très chronophages et leur automatisation ~~semble~~ aujourd'hui nécessaire.

L'automatisation de tâches de classification à l'aide de méthodes statistiques a reçu un intérêt particulier en océanographie depuis 20 ans. C'est particulièrement le cas pour le traitement d'images. On peut par exemple citer le dénombrement de coccolithes de Beaufort et Dolfus (2004) [1] à l'aide de réseaux de neurones peu profonds ou plus récemment les travaux de Dunker & al. (2019) [2], González & al. (2019) [3] ou enfin le Projet RAPP en cours de développement au CEREGE, qui reposent sur l'utilisation de réseaux de neurones profonds et pré-entraînés.

La reconnaissance automatisée des groupes d'individus à partir des courbes issues de la cytométrie en flux a été moins explorée à certaines exceptions notables comme Malkassian (2011) [4] qui classe les espèces du phytoplancton en plongeant ces courbes dans une base de Fourier et en calculant des distances entre ces dernières.

La présente thèse se propose quant à elle de dénombrer les cellules appartenant à chaque groupe fonctionnel du phytoplancton (et non pas à chaque espèce) à l'aide des courbes cytométriques. Classer puis dénombrer les cellules à l'aide des courbes cytométriques présente un réel avantage en termes de poids de stockage par rapport aux images. De plus, la résolution des appareils ne permet pas de ~~photographier~~, à l'heure actuelle, des cellules inférieures à 20 microns, dont les courbes cytométriques existent, elles, cependant.

Afin de classer ces courbes, un prétraitement est tout d'abord nécessaire afin de rendre l'ensemble des courbes comparables et de prendre en compte le fait que certains groupes sont très peu présents dans les données. Des méthodes récentes issues de l'apprentissage profond supervisé comme les réseaux de neurones convolutifs (CNN) [5] ou à mémoire (LSTM, GRU) [6] ont ensuite permis de réaliser la classification. Les données utilisées pour l'apprentissage et la prédiction sont issues de la campagne FUMSECK.

Il ressort des simulations que les réseaux de neurones « sans mémoires » et dont les premières couches sont des couches convolutives semblent plus efficaces pour classer les groupes fonctionnels. Le pourcentage de cellules bien classé est proche de 90 % lors de nos premières estimations. Il semble cependant que le réseau parvienne à prédire de manière précise les groupes les plus représentés mais que les performances sur les groupes moins représentés dans les données soient nettement moins satisfaisantes.

D'autres travaux sont nécessaires afin de confirmer ces résultats et d'obtenir de meilleures prédictions concernant les groupes sous-représentés. Les méthodes envisagées consistent à améliorer le ré-échantillonnage préalable à la classification ou à pénaliser davantage les erreurs de prédiction réalisées sur ces groupes. Enfin, certains travaux [8] semblent montrer que classer afin de dénombrer les individus des différents groupes, approche appelée *« Classify and Count »,* est souvent sous-optimal et que des méthodes de « *quantification* » pourraient être plus adaptées.

Remerciements

Les remerciements doivent être écrits ici

Références

[1] Beaufort L. et Dollfus. D., *Marine Micropaleontology 51.1-2 (2004) 57–73.*

[2] Dunker S., *Cytometry Part A 95.8 (2019):* 854-868.

[3] González P. et al, *Journal of Plankton Research* 41.4 (2019) 449-463.

[4] Malkassian, A. et al., *Cytometry part A 79.4 (2011)* 263-275.

[5] Krizhevsky, A. et al. *Advances in neural information processing systems.* (2012) 1097-1105.

[6] Krizhevsky, A. et al. *Conference on Empirical Methods in Natural Language Processing (EMNLP),* 2014, Doha, Qatar.

[7] Gonzalez et al. *Progress in Artificial Intelligence*, 6 (2017) 53–58.