Atelier commun LEFE-CYBER / ILICO / ODATIS sur

L'utilisation de l'IA pour analyse de données issues de séries longues

Paris mardi 4 juin 14h - mercredi 5 juin 16h30,

Sorbonne Université (Jussieu), salle de réunion IPSL, couloir 45-55, 2ème étage

Les organisateurs : Sabine Schmidt (EPOC, LEFE-CYBER, ODATIS/IR DATA TERRA), Alain Lefebvre (Ifremer, IR ILICO), Raphaëlle Sauzède (CNRS, IMEV, Argo-France) et Saïd Ouala (IMT-Atlantique)

En raison de la croissance exponentielle de la quantité de données disponibles en milieux marins et côtiers, les méthodes conventionnelles d'exploitation des séries temporelles ne suffisent plus pour extraire toute l'information d'une Observation intégrée, multi-variable, multi-source et multi-résolution. L'expertise humaine face à ces séries peut se trouver limitée face à la quantité d'information désormais disponible pour répondre aux principaux défis scientifiques. Cela explique l'intérêt croissant suscité pour les méthodes basées sur les méthodes d'apprentissage : le Machine Learning (ML) classique et le Deep Learning (DL). Ces méthodes sont compatibles avec le traitement et l'analyse de jeux de données conséquents, capables d'évaluer efficacement la pertinence des variables considérées et de saisir les corrélations non linéaires entre les attributs à l'aide de techniques de modélisation simple, parfois sans avoir à comprendre le fonctionnement du système (méthode non supervisée) et même sans devoir obligatoirement considérer le temps et/ou l'espace comme contraintes pour définir des états environnementaux, les classifier et les prévoir. Les méthodes ML et DL deviennent ainsi des techniques incontournables pour la qualification, le traitement, l'analyse, la classification (clustering) des données et bien entendu la prédiction. Toutefois, leur usage n'est pas encore généralisé et se limite dans la communauté à quelques experts.

L'objectif de cet atelier est de rassembler les utilisateurs des techniques d'apprentissage ML/DL, d'échanger sur les usages, les jeux de données auxquels elles peuvent s'appliquer (in situ, imagerie, satellite, modèle), les applications possibles, et les précautions d'usage. Il est ouvert aux collègues non utilisateurs qui souhaiteraient être informés sur ces méthodes. Il est coorganisé par Sabine Schmidt (EPOC, LEFE-CYBER, ODATIS/IR DATA TERRA), Alain Lefebvre (Ifremer, IR ILICO), Raphaëlle Sauzède (CNRS, IMEV, Argo-France) et Saïd Ouala (IMT-Atlantique)

Après des présentations d'ensemble et d'autres plus ciblées sur des cas d'usage lors de la 1ère demi-journée, le second jour sera consacré aux discussions. Cet atelier, mettant l'accent sur les discussions, devrait permettre de présenter aux communautés diverses applications de techniques d'apprentissage automatique et profond, favorisant ainsi l'établissement de nouvelles collaborations et l'émergence de nouveaux projets. Des recommandations sur les données IA-compatible pourront aussi être émises à l'intention du pôle de données océan ODATIS de l'IR Data Terra et de l'IR ILICO.

Ordre du jour

Le mardi 04 juin

13h30- 14h00 Accueil

14h00 – 14h20 Introduction de l'atelier dans un contexte de multiplication des sources de données (incluant brève présentation de l'AO LEFE CYBER et d'ODATIS) Sabine Schmidt

Brève présentation de l'IR ILICO Alain Lefebvre

14h20 – 15h00 Introduction à l'intelligence artificielle dans l'assimilation des données géophysiques– Saïd Ouala

15h00 – 17h30 Aperçu des usages de l'IA / exemples

Les orateurs, exposés et temps d'échange : les orateurs sont priés de présenter en 20 minutes max. leurs démarches et éventuellement en concluant sur les avantages / inconvénients / limites de la méthode

Elodie Martinez (IRD) Utilisation du machine learning pour reconstruire des séries longues passées et futures de biomasse phytoplanctonique dans l'océan global

Lefebvre A. (Ifremer), Halawi Ghosn Raed (Ifremer), Wacquet G (Ifremer) & Poisson Caillault Emilie (LISIC, ULCO) Utilisation du Machine Learning pour la caractérisation et la prédiction des blooms phytoplanctoniques (y compris les HAB) et des états environnementaux associés.

Guillaume Wacquet (Ifremer) & Lefebvre A. (Ifremer), Application du machine learning à l'imagerie du phytoplancton

Laurent Coppola (IMEV, visio) CANYON-MED : un outil pour la prédiction de variables biogéochimiques

Raphaëlle Sauzède (IMEV) Application du Machine Learning aux données BGC-Argo: Nouveaux produits opérationnels et applications (IMEV) QC (Quality Assurance / Quality Control)

Dimitry Khvorostyanov (LOCEAN), Nathalie Lefevre Using machine learning algorithms to characterise the carbon system variability drivers in the eastern tropical Atlantic

17h30 – 18h : choix des thèmes à privilégier pour la seconde ½ journée

Les suggestions de thèmes à aborder sont:

- Les données : minimum de données requis, le type de données (in situ, satellite, modélisation), format, mais aussi quid de l'aide à qualifier des données par exemple ;
- La philosophie : combiner des données disparates, régulariser des séries, alignement temporel, impact des lacunes de données, …
- Les systèmes d'alerte : faisabilité, quelles méthodes,
- Les stratégies adaptatives lors des campagnes en mer (sites, variables mesurées, fréquences de mesures)
- Support d'assimilation de données pour les jumeaux numériques : limite à cet usage, fréquence de la mise à jour des relations etc
- Emergence d'autres méthodes de traitement ?

Le mercredi 05 juin

9h – 12h30

9h – 9h30 Deloffre et al (Univ Rouen) Utilisation de l'IA pour le traitement des données du continuum Terre-Mer.

9h30 – 12h30 Discussion des thèmes retenus

14h – 15h30-16h00

14h-14h30 Sudre Joël et l'IR Data Terra Développement d'une IA pour le portail de découverte Gaia Data

14h30 – 16h Discuter des suites à donner à l'atelier : recoms possibles les besoins de développements éventuels (CES ODATIS, production d'une note par exemple) ; intérêt d'un projet LEFE-CYBER à l'AO2024 (pour tester/intercomparer des méthodes par exemple).

Bref compte-rendu de l'atelier IA

Au 4 juin, il y avait 62 inscrits à l'atelier, répartis de façon assez équivalente entre présentiel et visio, et utilisateurs et non utilisateurs des techniques d'apprentissage. Dans la pratique, il n'y a pas eu de pointage des présences, plusieurs collègues ne sont finalement pas venus en présentiel.

L'atelier a débuté par des interventions des organisateurs :introduction de l'atelier, présentations de l'AAP LEFE CYBER, de ODATIS et de l'IR Ilico et une présentation de l'usage de l'intelligence artificielle dans l'assimilation des données géophysiques.

L'atelier s'est poursuivi par des exposés de collègues illustrant la diversité des usages. Lors de l'inscription, les collègues pouvaient indiquer s'ils souhaitaient présenter leurs travaux ayant recours à l'IA. Il y a eu au total 7 propositions qui couvraient des usages variés depuis l'océan hauturier aux estuaires. Ces exposés sont détaillés dans le programme et les exposés disponibles

Il y a eu de nombreux échanges en lien avec ces exposés dont il n'est pas fait un compte-rendu puisque les exposés sont fournis. Toutefois, un constat commun est qu'il est impératif de formuler le problème que l'on veut résoudre.

Des questions ont été posées sur le choix de la méthode, les variables forçantes et la phase d'entrainement (les jeux de données nécessaires). Ce choix peut-il entrainer un biais selon ce choix dans la mesure où c'est l'usager qui choisit ? Il ressort une tentation des collègues non encore utilisateurs d'appréhender les usages possibles et de recourir à ces techniques IA.

Un sujet récurrent a concerné les jeux de données et la question de définir des données compatibles IA. Il est fait état d'un constat qu'avoir des données à disposition c'est bien, mais une suggestion est faite que les jeux de données d'entrainement comportent une colonne label en complément d'un code qualité, pour utiliser l'expertise des producteurs/experts pour identifier des processus (tempête, bloom, etc).

En fait les échanges autour données IA compatibles soulèvent des questions, comme mettre en place le moyen de sauvegarder une expertise sur les données pour garder une mémoire numérique sur les données. L'argument est que ce serait un gain temps sur l'identification des processus connus pour aller plus loin. Une interrogation porte sur le moyen de mettre en place une telle mémoire, le vocabulaire associé et le niveau de confiance qui y serait accordée. La préparation des données en vue de l'IA est une étape longue et il faudrait favoriser des rétroactions entre la base de données (producteurs) et les utilisateurs ultérieurs. Mais ce problème n'est pas spécifique à l'IA, de base les données requises pour l'IA doivent répondre aux critères FAIR (métadonnées riches et bien renseignées, format ouvert, vocabulaire).

Un autre questionnement portait sur les données reconstituées. L'IA peut permettre de remplir des lacunes de données mais il faut pouvoir isoler la donnée mesurée et la donnée simulée. Dans l'exemple d'ARGO, il y a les données brutes qualifiées et il y a des produits créés, qui ont des incertitudes associées. Il faut que les collègues soient conscients de ce qu'ils utilisent.

La considération des lacunes de données dépend aussi de la problématique et peut influencer le choix de la méthode puisque certaines méthodes n'acceptent pas forcement les trous de données. Souvent la crainte avec les lacunes de données est une sous-représentation d'évènements rates mais le choix des périodes d'entrainement peut aussi déboucher sur une surreprésentation.

Certains exposés ont ouvert des perspectives pour les réseaux d'observation comme COASTHF. En effet ces réseaux ont parfois des lacunes de données. Il pourrait être intéressant de boucher ces trous, comme cela a été présentés pour le port de Havre. Mais cela pose encore la question de la représentation des évènements extrêmes moins fréquents.

Le fait qu'environ la moitié des inscrits étaient des collègues curieux de l'IA mais non utilisateurs soulève la question de l'appropriation de la méthode. Tout le monde peut-il se servir des données et des scripts ? Une tentation serait de reprendre des scripts utilisés sur des mêmes types de données. Serait-il envisageable de mettre à disposition une boite à outils et des codes utilisables pour d'autres personnes ? Les avis sont partagés. Il existe des sites avec des formations à l'IA, ou bien il est possible de se rapprocher de collègues/équipes qui ont une expertise.

Discussion des suites à donner à l'atelier :

La proposition d'un Consortium d'Expertise Scientifique (CES ODATIS) a été abordée. L'atelier a montré l'intérêt des communautés scientifiques, mais il faudrait une réflexion plus avancée sur le périmètre de possible CES (focus sur les données, mise à disposition de boites à outils). A envisager à moyen terme.

Par contre l'atelier a initié une dynamique qu'il faut maintenir. La proposition est que le colloque EVOLECO (novembre 2024, Brest) dédie une session du ce thème.