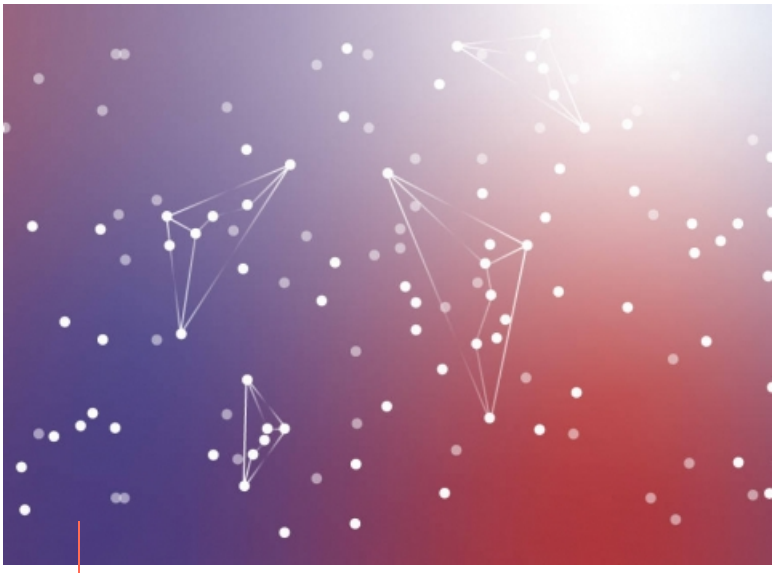
Intelligence artificielle pour la science des données et la cybersécurité

Les chercheurs dans cette thématique étudient de nouveaux modèles et de nouveaux algorithmes dans le domaine de l'intégration de données, de la classification et du (co)-clustering. Ils développent également des recherches théoriques et appliquées dans le domaine de la cybersécurité et de la gestion des ressources dans les systèmes cyber-physiques.



©CB / V. DEVILLARD

Référents scientifiques

Coordinateur : Mohamed Nadif

Responsables scientifiques :

- Severine Affeldt
- Lazhar Labiod
- Ahmed Mehaoua
- Osman Salem

Chiffres clés

0 post-doctorants -- 7 doctorans -- 6 thèses récemment soutenues

Présentation de la thematique

- L'apprentissage machine pour la science des données est un champ d'étude incontournable de l'intelligence artificielle. Il se décline sous différentes formes : non supervisée, semi-supervisée, supervisée et par renforcement. Bien que de multiples algorithmes, modèles et stratégies soient aujourd'hui disponibles, beaucoup de défis majeurs restent toujours d'actualité, et ceci dans des domaines aussi nombreux que variés.
- Les chercheurs de l'équipe se concentrent sur plusieurs questions liées à l'apprentissage notamment le (Co)-clustering et la réduction de la dimensionnalité (apprentissage non/semi supervisé) ainsi que la classification supervisée. Nos travaux reposent sur différentes approches telles que la factorisation matricielle, les modèles de mélange, les modèles des blocs latents, la décomposition spectrale et l'apprentissage profond. Notre objectif principal est de proposer des modèles et algorithmes innovants qui sont efficaces et facilement exploitables en pratique. Ainsi, les méthodes que nous proposons sont dédiées au traitement de données multi-sources et de différentes natures avec des applications dans divers domaines telles que l'analyse des données textuelles, le traitement automatique du langage naturel, la bioinformatique, le filtrage collaboratif, l'analyse de médiation et la cybersécurité. En outre, nous nous intéressons au domaine médical, en développant de nouvelles méthodes d'apprentissage machine et des logiciels conviviaux pouvant intégrer divers types de données omiques pour identifier les acteurs des maladies humaines complexes.
- Les chercheurs de l'équipe développent également des recherches théoriques et appliquées dans le domaine de la cybersécurité et de la gestion des ressources dans les systèmes cyber-physiques en particulier de détection d'anomalie pour les réseaux de capteurs de corps médicaux sans fil. Les contributions du groupe sont orientées sur la conception, l'optimisation et l'évaluation des performances des nouveaux protocoles, algorithmes, outils et modèles formels. Elles permettent, ainsi, de fournir une qualité, et une sécurité des communications et des données, dans les systèmes physiques de la santé de la prochaine génération telles que la détection des maladies chroniques (Ischémie, épilepsie, etc.).

Mots clés

Machine and Deep Learning ; Co-clustering ; Factorization ; Spectral Clustering ; Mixture models ; Attributed Network Embedding ; Mediation analysis ; Wireless Sensor Networks ; Internet of Medical Things ; Security and Anomaly detection ; Resource Optimization

Thèmes abordées

- Spectral clustering via ensemble deep autoencoder learning and evaluation on image data.
- Regularized bi-directional co-clustering for biomedical texts.
- Endotypes identified by cluster analysis in asthmatics and non-asthmatics and their clinical characteristics at follow-up: the case-control EGEA study.
- Unsupervised text mining for assessing and augmenting GWAS results.
- Real-time biomedical data analysis systems based on Machine Learning (ML) and Wireless Body Sensor Networks (WBAN),
- Sensor-based remote health monitoring, Sensor-based Human activity recognition, Sensor-based Ischemia and Epilepsy detection,
- Cybersecurity threats Detection using AI/ML, Blockchain-based Anomaly and threats detection for Internet of Things (IOT).

Applications

(Co)-clustering de documents textuels

- CIKM' 2021 : How to Leverage a Multi-layered Transformer Language Model for Text Clustering: an Ensemble Approach. (https://dl.acm.org/doi/abs/10.1145/3459637.3482121)
- SIGIR' 2021 : Regularized Dual-PPMI Co-clustering for Text Data (https://dl.acm.org/doi/abs/10.1145/3404835.3463065).
- CIKM' 2020: Ensemble block co-clustering: a unified framework for text data. (https://dl.acm.org/doi/10.1145/3340531.3412058)

(Co)-clustering d'images

- Spectral clustering via ensemble deep autoencoder learning. (https://www.sciencedirect.com/science/article/abs/pii/S0031320320303253)

Faits marquants

- Poisson degree corrected dynamic Stochastic Block Model (https://link.springer.com/article/10.1007/s11634-022-00492-9), P. Riverain, S. Fossier, M. Nadif, 2022
- Implicit consensus clustering (https://link.springer.com/article/10.1007/s10618-021-00788-y). R. Boutalbi, L. Labiod, M. Nadif.
- Regularized Bi-Directional Co-Clustering (https://link.springer.com/article/10.1007/s11222-021-10006-w), S. Affeldt, L. Labiod, M. Nadif, 2021.
- Unsupervised and self-supervised deep learning approaches for biomedical text mining, briefing in Bioinformatics (https://academic.oup.com/bib/article-abstract/22/2/1592/6132597), M. Nadif, F. Role, 2021.
- Man-in-the-Middle Attack Mitigation in Internet of Medical Things (https://ieeexplore.ieee.org/abstract/document/9456085), O. Salem, K. Alsubhi, A. Shaafi, M. Gheryani, A. Mehaoua, R. Boutaba, 2021,
- Markov Models for Anomaly Detection in Wireless Body Area Networks for Secure Health Monitoring (https://ieeexplore.ieee.org/abstract/document/9208756), O. Salem, K. Alsubhi, A. Mehaoua, R. Boutaba, 2021,

Softwares

- TensorClus (https://tensorclus.readthedocs.io) : A Python Library for Tensor (Co)-Clustering.
- CAEclust (https://github.com/MLDS-software/CAEclust) : A consensus of autoencoders representations for clustering
- CoClust (https://www.jstatsoft.org/article/view/v088i07) : A Python Package for Co-Clustering.

Projets

- Projet ANR GePhEx (https://sites.google.com/site/severineaffeldt/gephex/gephex-team) (S. Affeld, 2019) : Learning causal effects between phenome and exposome from large amounts of heterogeneous data in human complex diseases.
- Projet ANSES MOLDASTH (https://www.anses.fr/fr/system/files/APR2021_Liste_des_projets_finances.pdf) (R. Nadif, 2021) Moulds in dwellings, inflammation, immune response, and ASTHma endotypes in the CONSTANCES cohort.
- Projet Emergence Idex Spectrans (https://www.clillac-arp.univ-paris-diderot.fr/projets/spectrans) (M. Nadif, 2021). Specialised corpora and neural translation.
- Projet CDC Informatique (https://www.icdc.caissedesdepots.fr/). (M. Nadif) Détection d'anomalie et de contreverses en finance.
- Projet THALES (https://www.thalesgroup.com/fr) (M. Nadif, 2019), Hybridation d'algorithmes d'IA avec des connaissances métier pour le transport ferroviaire.
- Project SOPRA-STERIA (https://www.soprasteria.fr/) and AIRBUS-APSYS. Security of Industrial Internet of Things based on Blockchain.
- Project ORANGE LABS. Real-time Network Service Detection, Classification and Analysis from encrypted real-time traffic communications

Interactions avec les autres thématiques du Centre Borelli

Apprentissage statistique et analyse de données massives

Cette thématique s'occupe de l'inférence, de la modélisation prédictive et de l'optimisation séquentielle à partir de données complexes telles que des séries temporelles, des données fonctionnelles,...