AfIA

Association française
pour l'Intelligence Artificielle

CNIA

Conférence Nationale en Intelligence Artificielle

PFIA 2024

Table des matières

Éditorial

Conférence Nationale en Intelligence Artificielle

La Conférence Nationale en Intelligence Artificielle (CNIA), soutenue par le Conseil d'Administration de l'AFIA, s'adresse à l'ensemble de la communauté de recherche en IA. CNIA se veut un lieu privilégié pour faire connaître les dernières avancées en IA. Elle se veut aussi un forum destiné à renforcer les liens et les interactions entre les différentes sous-disciplines de l'IA et les disciplines faisant appel à l'IA. À ce titre, CNIA encourage les soumissions à la frontière entre sous-branches de l'IA, ainsi que les soumissions à la frontière de l'IA et d'autres disciplines.

Alors que l'IA se trouve aujourd'hui au cœur de nombreux développements, il est important d'avoir un forum qui réunisse l'ensemble des acteurs intéressés de près ou de loin par l'IA. L'objectif de CNIA est d'aborder à la fois les problématiques de recherche, les enjeux technologiques et les enjeux sociétaux liés à l'utilisation de l'IA, à travers l'ensemble des disciplines de l'IA :

- recherche heuristique et résolution de problèmes,
- incertitude et intelligence artificielle,
- logique, satisfiabilité et satisfaction de contraintes,
- apprentissage automatique,
- extraction, ingénierie et gestion des connaissances,
- représentation des connaissances et raisonnement,
- planification, contrôle,
- aide à la décision,
- causalité,
- agents autonomes et systèmes multi-agents,
- reconnaissance des formes et vision par ordinateur,
- traitement automatique des langues naturelles et de la parole, recherche d'information,
- interactions de systèmes d'IA avec l'humain,
- perception et robotique,
- IA et web,
- environnements informatiques d'apprentissage humain et d'apprentissage à distance,
- IA responsable, IA de confiance (incluant explicabilité, certification, équité, ...),
- éthique de l'IA,
- droit et IA,
- IA et société,
- IA dans divers domaines d'application comme la santé, l'environnement, l'énergie, le transport, la défense, l'agriculture, les matériaux, ...

Pour cette édition 2024 de CNIA, 20 articles ont été soumis. Chacun a été relu par 3 relecteurs membres du comité de programme. A l'issue de ce processus, 11 articles ont été retenus (soit un taux d'acceptation de 55%) pour être présentés lors de la conférence : 5 articles longs, 4 articles courts et 2 résumés en français d'articles présentés lors de conférences internationales. Les actes de CNIA regroupent ces articles en 5 sessions, dont les thèmes confirment l'importance actuelle des recherches sur l'apprentissage automatique en IA, y compris au sein de la communauté française. La définition de nouveaux modèles d'apprentissage automatique et leur optimisation pour des tâches spécifiques sont à la fois un terrain de recherche très fertile et une source de renouvellement et de d'avancées pour d'autres champs de l'IA, comme la logique, le traitement automatique du langage ou encore la représentation des connaissances, en particulier autour de la question de l'explicabilité de ces modèles et de la confiance à accorder à leurs résultats.

La conférence s'est déroulée sur 3 jours du 3 au 5 juillet 2024 suivant un programme découpé en 10 sessions. Parmi celles-ci, nous avons eu le plaisir et l'honneur d'accueillir deux conférenciers invités, que nous remercions chaleureusement ici. Chacun d'eux a abordé différentes facettes des enjeux des recherches sur les grands modèles de langue qui rendent aujourd'hui très visibles auprès du grand public les avancées des recherches en IA :

— Pierre Zweigenbaum (CNRS LISN et université Paris-Saclay) a abordé ces modèles sous l'angle de leur apport au traitement automatique du langage naturel (TAL) dans les documents de santé ; cette conférence est proposée à l'ensemble des participants de PFIA ;

— Eric Gaussier (LIG et Université Grenoble - Alpes) s'est focalisé sur la généralisation des réseaux de neurones en faisant un parallèle sur leur utilisation en TAL et en recherche d'information (RI) ; cette conférence invitée est propre à CNIA.

7 sessions ont été consacrées à des communications dont trois ont été organisées conjointement avec deux autres conférences de PFIA : IC et RJCIA. Outre les articles sélectionnés, ces sessions ont aussi permis d'exposer 8 communications de chercheurs de laboratoires français déjà présentées lors de conférences internationales (ces communications ne sont pas mentionnées dans les actes de CNIA).

Enfin, une session commune à CNIA, IC et RJCIA, a donné lieu à une table ronde pour alimenter les réflexions de la communauté de recherche en IA sur son avenir et ses évolutions, en éclairant les complémentarités entre approches symboliques, représentation des connaissances et apprentissage automatique. Merci à Cassia Trojahn (IRIT, Université Toulouse 2 Jean Jaurès), à Davide Buscaldi (LIPN, université Paris 13) ainsi qu'aux deux conférenciers invités d'avoir nourri le débat.

Je profite de cet éditorial pour remercier les membres du comité de programme pour leur précieux travail et pour la qualité de leurs relectures. Au nom du comité de programme, je remercie également l'ensemble des acteurs de la communauté francophone en IA qui ont contribué au succès de CNIA 2024, ainsi que le comité d'organisation de la plate-forme PFIA 2024 qui a été particulièrement efficace à toutes les étapes de l'organisation de la conférence et nous a simplifié la tâche au maximum. Enfin, tout le comité adresse ses plus vifs remerciements à Thomas Guyet pour son engagement et son soutien bienveillant et sans faille pour gérer la plate-forme au nom du bureau de l'AFIA.

Nathalie Aussenac-Gilles

Comité de programme

Présidence

— Nathalie Aussenac-Gilles, Université de Toulouse, CNRS, IRIT.

Membres

— Jérome Azé, Université de Montpellier, LIRMM ;
— Isabelle Bloch, Sorbonne Université, LTCI ;
— Olivier Boissier, Mines Saint-Etienne, LIMOS ;
— Robert Bossy, INRAE Centre de Jouy en Josas, MaIAGE ;
— Armelle Brun, Université de Lorraine, LORIA ;
— Cécile Capponi, Université d'Aix-Marseille, LIS ;
— Sylvie Coste-Marquis, Université D'Artois, CRIL ;
— Benjamin Dalmas, Centre de Recherche Informatique de Montréal ;
— Yves Demazeau, CNRS, LIG ;
— Sébastien Destercke, HDS, Université Technologique de Compiègnes, Heudiasyc ;
— Arnaud Doniec, IMT Lille Douai ;
— Jérôme Euzenat, INRIA Alpes ;
— Jean-Gabriel Ganascia, Sorbonne Université, LIP6 ;
— Eric Gaussier, Université Grenoble Alpes et IUF, LIG ;
— Guillaume Gravier, CNRS, IRISA ;
— Nathalie Hernandez, Université de Toulouse, UT2, IRIT ;
— Andréas Herzig, Université de Toulouse, CNRS, IRIT ;
— Céline Hudelot, Ecole Centrale Paris ;
— Camille Kurtz, Université Paris Cité ;
— Nicolas Lachiche, Université de Strasbourg ;
— Frédérique Laforest, INSA Lyon, LIRIS ;
— Florence Le Ber, École Nationale du Génie de l'Eau et de l'Environnement de Strasbourg, iCUBE ;
— Philippe Lenca, IMT Atlantique ;
— Marie-Jeanne Lesot, Sorbonne Université, LIP6 ;
— Pascal Poncelet, Université de Montpellier, LIRMM ;
— Catherine Roussey, INRAE Centre Occitanie-Montpellier, MISTEA ;
— Pascale Sébillot, INSA Rennes, IRISA ;
— Nazha Selmaoui-Folcher, Université de la Nouvelle Calédonie, ISEA ;
— Laurent Vercouter, INSA Rouen Normandie, LITIS ;
— Bruno Zanuttini, Université de Caen Normandie, GREYC.

Session 1 : Apprentissage automatique pour la génération et l'analyse de musique et d'images

Génération d'images de la Méiofaune à l'aide de StyleGAN2 : Cas des Copepoda

A. Martinel[1], A. Benzinou[1], K. Nasreddine[1], V. Foulon[1], C. Borremans[2], D. Zeppilli[2]

[1] ENIB, UMR CNRS 6285 LabSTICC, 29238 Brest, France
[2] Ifremer, 29280 Plouzané, France

Résumé

Nous explorons différentes approches hiérarchiques de transfert d'apprentissage d'un réseau antagoniste génératif StyleGAN2 afin de synthétiser des images de Copepoda. Il s'agit d'un des groupes les plus abondants de la faune aquatique, possédant peu d'images disponibles publiquement. Ces animaux sont de formidables bio-indicateurs de la pollution ou des changements environnementaux d'un milieu. Deux schémas d'apprentissage sont proposés. Le premier consiste à pré-entraîner le réseau avec les données d'un autre spécimen de même rang taxonomique et de faire un transfert d'apprentissage sur les données de l'animal étudié. Le deuxième consiste à pré-entraîner le réseau pour capturer les caractéristiques communes aux spécimens d'un rang taxonomique supérieur, pour enfin affiner le modèle au rang taxonomique inférieur souhaité. Ces méthodes visent à profiter des relations qui lient différents rangs taxonomiques. Les modèles obtenus sont ensuite évalués à l'aide des métriques FID et KID. Les images générées sont prometteuses, montrant des caractéristiques morphologiques typiques des copépodes. Ces données pourront ensuite être utilisées pour la formation de futurs taxonomistes et pour le développement de classifieurs de ces animaux, modèles qui nécessitent un grand nombre d'images pour leur entraînement.

Abstract

We explore two StyleGAN2 hierarchical transfer learning approaches in order to generate synthetic images of Copepoda animals. This is one of the most abundant groups in the aquatic fauna, yet only a few publicly available images are available. These animals are formidable bio-indicators of environmental changes or pollution of an habitat. Two learning approaches are proposed. The first is to pre-train the network with a dataset of another specimen of the same taxonomic rank and then fine tune it with the dataset of the studied animal. The second is to pre-train the network to capture features common to specimens of a higher taxonomic rank, finally refining the model to the desired lower taxonomic rank. Both methods aim to take advantage of the relationships that link different taxonomic ranks. The resulting models are then evaluated using the FID and KID metrics. The generated images are promising, showing typical morphological features of Copepods. These data can then be used to train future taxonomists and develop classifiers, models that require a large number of images for training.

Keywords

Apprentissage automatique, Réseaux adversoriels génératifs, StyleGAN2, Meiofauna

1 Introduction

La meiofaune désigne l'ensemble des espèces animales d'une taille comprise entre $20\mu m$ et $1mm$, en particulier les espèces vivant dans les sédiments au fond de l'eau sont dites benthiques. Les mots meiobenthos et meiofaune sont toutefois utilisés de façon ambigüe. Ces espèces jouent un rôle primordial au sein de l'écosystème marin profond, à la fois en tant que producteurs et consommateurs [14]. Elles réagissent plus rapidement que la macrofaune aux différentes formes de stress auxquelles elles sont exposées, faisant d'elles d'excellents bio-indicateurs des changements écologiques, de la pollution du milieu ou encore du dérèglement climatique : chalutage, décharge industrielle et agricole, pollution plastique, pollution aux métaux lourds et hydrocarbures, rejet de déchets nucléaires, déchets de munitions, exploitation minérale marine, acidification et désoxygénation des océans, etc. Avec 11 290 espèces identifiées [4], le groupe des Copepoda est le deuxième groupe le plus représenté de la méiofaune et également l'un des groupes les plus importants parmi les organismes planctoniques de cette taille. Les copépodes sont des animaux benthiques et pélagiques (vivant dans les sédiments et la colomne d'eau) dont la taille varie entre $250\mu m$ et $3cm$. Comme pour d'autres groupes de la méiofaune, le nombre d'espèces inconnues est très important et les estimations varient énormément : entre 30 125 et 450 000 espèces de copépodes encore inconnues selon des études [12, 1]. La figure 1 montre quelques exemples d'images de copépodes.

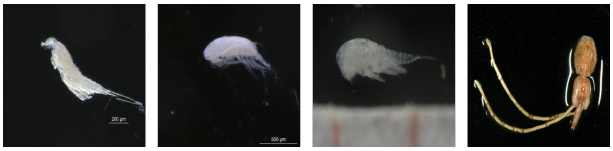

Figure 1: Exemples d'images de copépodes. De gauche à droite, un animal de l'ordre des Calanoida, Cyclopoida, Harpacticoida et Siphonostomatoida.

Néanmoins, l'identification taxonomique de la méiofaune demeure une tâche difficile. Elle repose sur l'utilisation de clefs taxonomiques et nécessite l'expertise de taxonomistes spécialistes des espèces étudiées, requérant plusieurs semaines de travail pour un échantillon de quelques centaines d'individus. Ce travail est rendu plus ardu par le manque de main d'œuvre, faisant du développement d'outils d'intelligence artificielle un enjeu majeur pour l'étude de la méiofaune. Cependant, ce développement est entravé par le manque de données. Motivés par les récentes améliorations apportées aux réseaux antagonistes génératifs (GANs) de haute définition entrainés sur de petits ensembles de données, nous étudions dans ce travail l'utilisation de réseaux neuronaux artificiels pour la génération de données synthétiques de copépodes.

Introduits pour la première fois en 2014, les GANs [5] sont devenus très populaires dans divers domaines de la génération de données, tels que la génération d'images, la traduction d'image à image ou de texte à image, l'augmentation des données ou encore le débruitage. Les GANs se composent de deux réseaux de neurones artificiels concurrents. Le premier, appelé générateur, crée à partir d'un bruit aléatoire de fausses données qui doivent tromper le second, appelé discriminateur. Ce dernier est entraîné à classifier entre données réelles et données générées. À la suite de la première architecture proposée [5], de nombreux nouveaux modèles ont vu le jour améliorant la qualité et la résolution des images générées. Récemment, deux GANs ont considérablement amélioré la qualité des images synthétiques ; les réseaux BigGANs [3] et StyleGANs [8] sont capables de générer des images de haute résolution, typiquement 512x512 pixels voir même 1024x1024 pixels. Le principal inconvénient des BigGANs est la taille importante de leur réseau, qui nécessite un très grand nombre d'images pour l'entrainement. Les améliorations récentes sur la seconde version du StyleGAN, à savoir le StyleGAN2 [9], pour les jeux de données limités nous ont conduit à choisir ce réseau pour base de notre travail. Nous proposons ici deux approches hiérarchiques de transfert d'apprentissage pour son entrainement à générer des images de Copépodes.

Le modèle, ses améliorations et les méthodes d'apprentissage que nous proposons sont expliqués plus en détail dans la section suivante. En sections 3 et 4, nous décrivons les données d'entrainement et les critères d'évaluation. Enfin, nous présentons et discutons les résultats obtenus dans la section 5.

2 Méthode proposée

Introduit en 2014, le DCGAN [10] est l'une des architectures convolutionnelles les plus populaires ; ce réseau est moins coûteux en terme de calcul et moins complexe à entraîner que d'autres GANs qui lui ont succédé. La particularité du StyleGAN est l'introduction de vecteurs de style : au lieu d'alimenter directement le générateur avec un vecteur de bruit gaussien, un réseau dense appelé réseau de mapping le transforme au préalable en un vecteur d'espace latent intermédiaire. Le vecteur latent obtenu est transformé à son tour via des transformations affines en un vecteur de style **A**. Ce dernier est ensuite introduit dans chaque bloc du générateur. Une seconde version, StyleGAN2, améliore son prédécesseur, en augmentant ses performances et en évitant l'apparition d'artefacts. Les blocs de base du générateur de StyleGAN2 sont illustrés dans la Figure 2, le vecteur de style **A** est issu du réseau de mapping, dans notre cas un Perceptron de deux couches. Les variations stochastiques de l'image sont fournies par l'injection parallèle d'un vecteur de bruit **B**. La modulation met à l'échelle les poids des caractéristiques d'entrée en fonction du style **A** : $w'_{ijkl} = s_i \cdot w_{ijkl}$, où w et w' sont respectivement les poids originaux et modulés, k et l sont les indices spatiaux, j est l'indice de la carte de caractéristique de sortie, et s_i est le style correspondant à la carte de caractéristiques d'entrée de rang i. Après cette opération, les activations de sortie ont un écart type de :

$$\sigma_j = \sqrt{\sum_{i,k,l} {w'_{ijkl}}^2} \tag{1}$$

Une opération de démodulation est appliquée :

$$w''_{ijkl} = \frac{w'_{ijkl}}{\sqrt{\sum_{i,k,l} {w'_{ijkl}}^2 + \epsilon}} \tag{2}$$

où ϵ est une constante qui permet d'éviter des instabilités numériques.

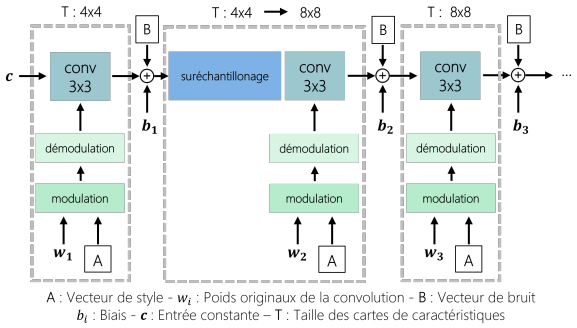

Figure 2: Schéma des trois premiers blocs du générateur du StyleGAN2. Les blocs suivants se succèdent jusqu'à atteindre la taille de l'image de sortie.

Enfin, nous proposons d'utiliser la limite adaptative d'augmentation de données ADA (Adaptive Discriminator Augmentation) [7] et la régularisation LeCam R_{LC} [13]:

$$R_{LC} = \mathop{\mathbb{E}}_{x\sim\tau}[\|D(x) - \alpha_F\|^2] + \mathop{\mathbb{E}}_{z\sim p_z}[\|D(G(z) - \alpha_R\|^2] \tag{3}$$

où τ est le jeu de données d'apprentissage, p_z est la distribution préalable, G et D sont respectivement le générateur et le discriminateur, et enfin α_F et α_R sont deux moyennes mobiles exponentielles. Ces méthodes permettent de stabiliser l'apprentissage dans le cas de jeux de données limités.

Pour l'entrainement, nous proposons deux protocoles afin d'obtenir des modèles capables de génération d'images synthétiques selon le rang taxonomique souhaité. Notre première approche vise à bénéficier des connaissances apprises

sur un autre spécimen que celui étudié. Elle consiste à entraîner d'abord le réseau avec un jeu de données d'un autre spécimen de même rang taxonomique pour lequel nous avons plus d'images et ensuite avec le jeu de données de l'animal étudié. Cela permet d'apprendre les caractéristiques communes d'un spécimen de même rang taxonomique avant d'utiliser les images de l'animal à synthétiser pour spécifier ses propres caractéristiques. Par la suite, nous notons cette méthode "méthode latérale". Nous proposons une seconde approche, permettant au réseau de capturer les caractéristiques communes aux spécimens d'un rang taxonomique supérieur, pour enfin affiner le modèle au rang taxonomique inférieur souhaité. Cette méthode hiérarchique est appelée "méthode descendante", ainsi par exemple nous entrainons nos modèles avec l'ensemble des images de Copepodes avant d'affiner en n'utilisant seulement les images de Calanoides.

3 Données d'entrainement

Les données utilisées dans cette étude proviennent du jeu de données BOLD (The Barcode of Life Data System) [11]. Les images sont triées pour ne garder que les exemples de taille supérieure à 256 x 256 pixels et représentant l'animal en entier. Les images sont ensuite redimensionnées à 256 par 256 pixels. Les copépodes sont identifiés selon leur ordre taxonomique. L'ensemble de données est composé de 1969 Copepoda, dont 902 Calanoida, 766 Cyclopoida, 150 Harpacticoida et 151 Siphonostomatoida.

4 Métriques de performance

Pour évaluer la qualité des images générées, nous utilisons les mesures FID (Fréchet Inception Distance) [6] et KID (Kernel Inception Distance) [2]. Le FID mesure la différence entre les distributions statistiques des caractéristiques, générées à l'aide d'un modèle Inception pré-entraîné, des images synthétiques et réelles. Dans le cas du FID-50k, les caractéristiques de l'ensemble de données réelles sont comparées à l'aide de la distance de Fréchet aux caractéristiques de 50 000 images générées. Afin de valider nos méthodes d'entrainement, nous ajoutons une deuxième mesure à nos tests de méthodes d'apprentissage. En effet, Keras et al. [7] ont montré que le FID n'était pas une métrique idéale dans le cas de jeux de données limités. Dans ce cas, le KID qui est une mesure non biaisée est plus approprié pour juger de la qualité des images générées. Le KID est calculé à l'aide de l'écart moyen maximal (MMD) des caractéristiques générées (de la même manière que le FID), des images synthétiques et réelles. Les images générées de meilleure qualité présentent des valeurs FID et KID plus faibles.

5 Résultats et discussion

Afin de valider les performances des méthodes proposées, nous avons également entraîné et évalué un modèle DCGAN avec le même jeu de données Calanoida. Nous présentons dans le Tableau 1 les résultats quantitatifs des modèles DCGAN et StyleGAN2. L'architecture DCGAN a du mal à générer des images suffisamment réalistes pour s'apparenter à des images réelles. Le StyleGAN2 en revanche obtient de bien meilleurs résultats. Ayant validé l'avantage du StyleGAN2 par rapport au DCGAN, nous expérimentons la limite adaptative d'augmentation de données ADA et la régularisation LeCam qui améliorent toutes deux ses performances.

Table 1: Comparaison des performances des modèles entrainés avec le sous-ensemble des données Calanoida.

Model	**Fid50k** $\downarrow$
DCGAN [10]	239.240
StyleGAN2 [9]	137.426
StyleGAN2 + ADA [7]	44.774
StyleGAN2 + ADA + R_{LC} [13]	**39.341**

Figure 3: Images synthétiques de copépodes générées par la méthode hiérarchique descendante. $1^{ère}$ ligne : Calanoida, $2^{ème}$: Cyclopoida, $3^{ème}$: Harpacticoida et $4^{ème}$: Siphonostomatoida.

Dans le tableau 2, nous présentons les mesures de qualité des trois méthodes. Sans surprise, plus nous avons de données, meilleure est la qualité des images générées ; les modèles entraînés sur le jeu de données Calanoida montrent ainsi les meilleurs résultats. Nous observons que pour chaque spécimen, l'apprentissage par transfert descendant depuis les copépodes permet au modèle de générer des données de meilleure qualité. Cette approche hiérarchique de l'apprentissage par transfert, en commençant par l'entraînement au rang taxonomique plus général (ici Copepoda), puis en affinant le modèle au rang taxonomique plus spécifique semble être la méthode la plus efficace. Notons les différences de perception de qualité d'image des deux métriques utilisées ; en effet, lorsque les modèles entraînés sur les Cyclopoida et les Harpacticoida ont respectivement des KID proches, leurs FID respectifs sont très éloignés. Cette différence est selon nous due aux différences de taille et de diversité des deux jeux de données. En effet le jeu de

Table 2: Comparaison des performances des méthodes d'entrainement utilisées (directe, latérale et descendante). Pour chaque ordre nous entrainons sur son jeu de données, un modèle sans transfert avec les poids initialisés de manière aléatoire, un modèle pré-entrainé sur les données de Calanoida et un modèle pré-entrainé sur les données regroupées de Copepoda.

	Méthode directe		**Méthode latérale**		**Méthode descendante**	
Ordre	Fid50k ↓	Kid50k ↓ ($\times 10^2$)	Fid50k ↓	Kid50k ↓ ($\times 10^2$)	Fid50k ↓	Kid50k ↓ ($\times 10^2$)
Calanoida	39.341	1.204	X	X	**36.529**	**0.931**
Cyclopoida	60.401	2.327	50.504	1.545	**47.017**	**1.215**
Harpacticoida	119.298	2.323	114.288	2.070	**108.533**	**1.503**
Siphonostomatoida	111.865	1.611	107.610	1.549	**96.995**	**0.990**

données des Harpacticoida est beaucoup plus petit (150 images) que celui des Cyclopoida (766 images) et contient des images moins diversifiées (posture, microscope utilisé, etc). Ces mesures de performance ne quantifient pas forcément l'intégralité des traits caractéristiques de chaque spécimen. Afin de compléter ces résultats et valider la capacité de nos images à rendre compte des différences entre les ordres, nous avons présenté nos images à des experts mondiaux des Copépodes. Ces derniers ont confirmé la qualité de nos images. Des exemples d'images générées par nos modèles (entraînés via la méthode descendante) sont présentés dans la Figure 3. Les images générées présentent des caractéristiques complexes comme les réglettes, des réflexions lumineuses ou même des œufs.

6 Conclusion et perspectives

Nous avons proposé deux approches d'apprentissage par transfert hiérarchique pour la génération d'images de copépodes à l'aide d'un réseau StyleGAN2. Nous obtenons des résultats très prometteurs, à l'aide de nos deux stratégies les modèles utilisés ont réussi à générer des images d'animaux qui ont été identifiés comme des copépodes par des taxonomistes. Ces approches peuvent être utilisées sur d'autres espèces de la méiofaune afin de générer des données synthétiques de spécimens pour lesquels peu de données sont disponibles. Ces données synthétiques pourraient constituer un nouvel outil pour former des taxonomistes ou être utilisées pour l'apprentissage de réseaux classifieurs. Une lecture par des systématiciens taxonomistes a permis de valider la qualité des images synthétisées d'un point de vue taxonomique. En perspective, nous envisageons d'introduire de l'information de haut niveau dans les premières couches du réseau, comme les traits fonctionnels morphologiques systématiquement utilisés par les taxonomistes ainsi que de tester la troisième version du StyleGAN, censée améliorer la représentation interne des images.

References

[1] W. Appeltans, S. T. Ahyong, G. Anderson, M. V. Angel, T. Artois, N. Bailly, et al. The magnitude of global marine species diversity. *Current Biology*, 22(23):2189–2202, 2012.

[2] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. *ICLR*, 2018.

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096, 2018.

[4] M. J. Costello and C. Chaudhary. Marine biodiversity, biogeography, deep-sea gradients, and conservation. *Current Biology*, 27(11):R511–R527, 2017.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al. Generative adversarial networks. *NIPS*, page 2672–2680, 2014.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[7] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12104–12114. Curran Associates, Inc., 2020.

[8] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018.

[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. *CoRR*, abs/1912.04958, 2019.

[10] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2014.

[11] S. Ratnasingham and P. Hebert. Bold: The barcode of life data system (www.barcodinglife.org). *Molecular ecology notes*, 7:355–364, 06 2007.

[12] S. Seifried. Phylogeny of harpacticoida (copepoda): Revision of 'maxillipedasphalea' and exanechentera. *Cuvillier Verlag, Göttingen Germany Ph.D. thesis from 2002*, pages :1–259., 2003. The World Of Copepods (T. Chad Walter).

[13] H. Tseng, L. Jiang, C. Liu, M. Yang, and W. Yang. Regularizing generative adversarial networks under limited data. *CoRR*, abs/2104.03310, 2021.

[14] D. Zeppilli, J. Sarrazin, D. Leduc, P. Martinez Arbizu, D. Fontaneto, C. Fontanier, et al. Is the meiofauna a good indicator for climate change and anthropogenic impacts? *Marine Biodiversity*, 45, 09 2015.

Analyse de l'encodeur pour la segmentation d'une base de données hétérogène de photographies de plaies chroniques avec peu d'annotations

G. PICAUD[1,3], M. CHAUMONT[1,2], G. SUBSOL[1], L. TEOT[3]

[1] LIRMM, équipe ICAR, Univ. Montpellier, CNRS, Montpellier, France
[2] Univ. Nîmes Place Gabriel Péri, 30000 Nîmes Cedex 01, France [3] Cicat-Occitanie, Montpellier, France

{guillaume.picaud, marc.chaumont, gerard.subsol}@lirmm.fr, l-teot@chu-montpellier.fr

Résumé

La segmentation est cruciale en imagerie médicale mais l'obtention de données annotées en quantité suffisante est difficile, limitant le développement de modèles d'apprentissage profond performants. Les stratégies d'apprentissage auto-supervisé (SSL) offrent une solution prometteuse pour pallier ce manque d'annotation. L'une d'entre elles, Dinov2 pour Distillation with NO labels a permis l'élaboration de l'immense base de données LVD-142M ainsi que l'entraînement d'encodeurs, aujourd'hui en accès libre. Cependant, les images cliniques ne sont pas nécessairement bien représentées dans LVD-142M. Dans cet article, nous comparons différentes méthodes d'initialisation d'encodeurs pour la segmentation de photographies cliniques dans un contexte de manque d'annotation.

Mots-clés

Apprentissage auto-supervisé, segmentation, Dinov2, images cliniques.

Abstract

Segmentation task is crucial in medical imaging, but obtaining a sufficient quantity of annotated data is challenging, limiting the development of high-performing deep learning models. Self-supervised learning (SSL) strategies offer a promising solution to address this lack of annotation. One such strategy, Dinov2 for Distillation with NO labels, enabled the creation of the vast LVD-142M database and also the training of encoders, whose weights are now freely accessible. However, clinical images may not be well represented in LVD-142M. In this article, we evaluate the benefits of different encoder initialization methods for segmentation in a context of scarce annotated clinical data.

Keywords

Self supervised learning, segmentation, Dinov2, weight initialization, clinical images

1 Introduction

La Haute Autorité de Santé définit les plaies chroniques comme des lésions n'ayant pas atteint une cicatrisation complète après 4 à 6 semaines d'évolution. De multiples facteurs peuvent favoriser leurs apparitions au sein de populations à risque comprenant les personnes âgées, les diabétiques ainsi que les personnes à mobilité réduite. Elles posent un problème socio-économique majeur avec des conséquences sévères pour l'individu pouvant aller de l'amputation au décès du patient. L'assurance maladie a estimé à plus d'un milliard d'euros la seule gestion des escarres et ulcères à domicile pour l'année 2011. Leur prévalence est en constante augmentation, notamment en raison du vieillissement de la population.

Le "Réseau Cicat-Occitanie" fournit une assistance à la prise en charge des plaies chroniques par le biais de téléconsultation afin de mettre en contact des experts avec les équipes médicales de proximité. En plus de 20 ans d'expérience, cette initiative a généré une base de données considérable de plus de 133 000 images photographiques de plaies chroniques de tout type (escarre, ulcère, plaie du pied diabétique, etc...). Cette base de données représente une ressource précieuse mais sous-exploitée en raison du manque de standardisation du protocole d'acquisition et d'annotations disponibles. Des exemples provenant de cette base de données sont visibles avec la figure 1.

La segmentation revêt une importance cruciale dans la prise en charges des plaies chroniques car la réalisation de leurs calques tout au long du parcours de soin aide le corps médical à évaluer l'efficacité des traitements choisis et ainsi valider ou réfuter la pertinence du diagnostic établi. Cependant, le détourage manuel est une tâche complexe entraînant des écarts mesurables tant entre les annotateurs qu'entre les différentes propositions d'un même annotateur, même lorsque ces derniers sont des experts. Par ailleurs, la base de données Cicat-Occitanie est caractérisée par la diversité du matériel d'acquisition (smartphones utilisés), des scènes (domiciles différents avec des variations dans l'éclairage, la prise de vue, l'arrière-plan, la distance entre le smartphone et la plaie) et des plaies (de nature et de localisation variées), ce qui complique la tâche.

Les méthodes par apprentissage profond représentent aujourd'hui l'approche privilégiée pour la segmentation des plaies chroniques. Des compétitions internationales comme

le DFUC pour Diabetic Foot Ulcer Challenge [1] rendent disponibles des bases de données de plusieurs milliers d'images acquises en conditions hospitalières et annotées en segmentation par des experts. Cette compétition a d'ailleurs mis en évidence les performances du modèle HardNet-MSEG [6]. Toutefois, l'absence d'initiative similaire pour des images hétérogènes dites "into the wild", limite aujourd'hui le développement d'approches supervisées suffisamment robustes face à la diversité des cas cliniques.

Pour surmonter cet obstacle, l'apprentissage auto-supervisé (self-supervised Learning, SSL) apparaît comme une piste prometteuse car il permet aux réseaux de neurones d'apprendre à représenter plus efficacement les images sans nécessiter de supervision humaine. En particulier, la méthodologie DINO pour Distillation with No labels [1, 7] a mené à l'élaboration de l'immense base de données LVD-142M à partir de laquelle l'encodeur ViT (Vision Transformer) a été entraîné [3]. Cependant, cette base de données générique ne représente pas bien les images cliniques et l'état de l'art manque de références quant aux bénéfices apportées par son utilisation dans ce domaine spécifique.

Cet article vise à explorer, dans un contexte clinique spécifique de segmentation de plaies chroniques avec peu d'annotations, l'intérêt de l'encodeur générique ViT préentraîné avec DINO sur LVD-142M face à l'encodeur HardNet-MSEG, plus léger et initialisé aléatoirement. Nous nous intéressons également à l'effet que produit un préentraînement SSL DINO sur les données cibles effectué avant la tâche finale de segmentation. Enfin, nous mesurons l'impact de la quantité de données disponible sur les performances des différents scénarios d'entraînements proposés.

FIGURE 1 – 4 exemples illustrant la diversité des photographies de plaies chroniques en terme de localisation, de nature et de conditions d'acquisition.

2 Etat de l'art

Le SSL est une approche où un encodeur est entraîné durant une tâche dite prétexte qui, au lieu d'utiliser des annotations humaines, se fonde sur des labels générés automatiquement à partir de la donnée elle-même [8, 11] disponible en grande quantité. Suite à ce préentraînement, les poids de l'encodeur servent d'initialisation pour l'apprentissage sur une tâche finale. Parmi l'ensemble des méthodes SSL présente dans l'état de l'art, nous nous intéressons ici à l'approche discriminative illustrée par la tâche prétexte DINO proposée par META.

DINO utilise 2 encodeurs d'architecture identique, dont l'un est appelé élève et l'autre enseignant. Pour la tâche prétexte, une stratégie multi-crop est appliquée à l'image d'entrée : 2 vues "globales", dont la surface représente au moins la moitié de l'image d'origine, et n vues "locales", ayant une surface inférieure à 50 %, sont générées. L'encodeur enseignant recevra les 2 vues "globales" tandis que l'élève verra toutes les vues. Chacune de ces vues est augmentée différemment à l'aide de transformations spatiales et colorimétriques. Lors de l'apprentissage, les deux encodeurs produisent une représentation de ces vues qui seront transmises à leur tête de projection respective consistant à une suite de couches linéaires (MLP). Les représentations sont alors comparées par entropie croisée et les poids de l'élève sont mis à jour par rétropropagation du gradient. Les poids de l'enseignant sont eux mis à jour via une moyenne mobile exponentielle à partir de ceux de l'élève.

A l'aide de la base de données soigneusement assemblée LVD-142M, SSL DINO a permis l'entraînement d'encodeurs ViT de différentes échelles (small 21 M, large 307 M, giant 1100 M de paramètres) voir [2]. Les Transformers sont des architectures imposantes, gourmandes en ressources informatiques et ne se démarquent des approches convolutives que lorsque les bases de données sont de très grandes tailles.

HardNet, pour Harmonic Densely Connected Network [2], est une architecture convolutive améliorée de l'architecture DenseNet [4] dont le but est de réduire le temps d'inférence sans réduire les performances de l'encodeur. Pour ce faire, le nombre et la position des connections résiduelles au sein des blocs de convolutifs ont été modifiés. Dans le cadre de la compétition DFUC2022 [5], cet encodeur a été amélioré et connecté à un décodeur de segmentation appelé Lawin pour Large window attention [10]. Cette proposition nommée HarDNet-MSEG a atteint la première place de la compétition de segmentation, rendant de facto cette architecture intéressante dans l'analyse des plaies chroniques.

3 Préparation des bases de données

Un Faster RCNN [9] modifié a été entraîné comme détecteur de plaies via les données de la compétition DFUC2020 [3]. Il a été appliqué sur la base de données du réseau Cicat-Occitanie rassemblant plus de 133 000 images. Seules les images n'ayant qu'une seule plaie prédite sont conservées, soit 88 727 images venant constituer la base de données B_1. Nous sélectionnons alors aléatoirement 400 images dans B_1 afin que deux experts les annotent en segmentation manuellement à l'aide de l'outil labelme [4] constituant ainsi B_2. L'élaboration de B_1 et B_2 est illustrée dans la figure 2.

B_1 est découpée en 3 catégories dédiées aux préentraînements SSL qui utiliseront le même protocole que DINO : l'entraînement, la validation et le test avec un ratio respectif de 70%, 20%, 10%.

Les images B_2 sont issues de la catégorie test de B_1 et sont utilisées durant la tâche finale de segmentation. Elles sont réparties en 5 folds de division 70%, 10%, 20%. Afin d'évaluer l'impact de la quantité de données

1. https ://dfuc2022.grand-challenge.org/
2. https ://github.com/facebookresearch/dinov2
3. https ://dfu2020.grand-challenge.org/
4. https ://github.com/labelmeai/labelme

d'entraînement, 3 copies de ces 5 folds ont été réalisées. Chaque copie se voit ôter un certain nombre d'images d'entraînement choisies aléatoirement tandis que les parties validation et test restent inchangées. Finalement, nous obtenons 3 groupes de 5 folds où le pourcentage de données d'entraînement est respectivement de 70%, 50% et 25% de B_2.

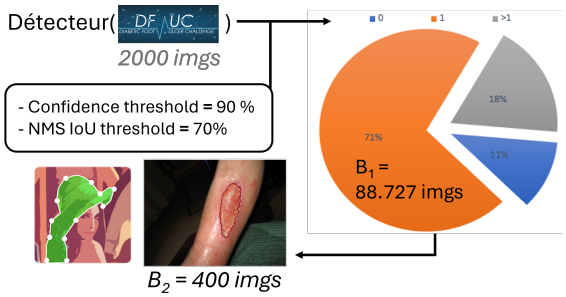

FIGURE 2 – Protocole de filtrage de la base de données Cicat-Occitanie : B_1 est réservée au SSL et B_2 à la segmentation.

4 Expériences

4.1 Scénarios d'entraînements

Concernant le choix des encodeurs, nous avons choisi les configurations ViT**s**14_reg (21 M) et ViT**l**14_reg (307 M) afin d'observer l'effet du changement d'échelle de la taille de l'encodeur. Leurs poids initiaux sont ceux issus de l'article Dinov2 [7]. Ces deux encodeurs sont comparés avec celui issu de HardNet-MSEG (3 M) dont l'initialisation est aléatoire.

La figure 3 résume les scénarios d'entraînement évalués. Les encodeurs peuvent être préentraînés ou non via la méthode SSL DINO sur B_1. Durant la tâche finale, les poids des encodeurs sont soit figés, soit optimisés au travers d'une stratégie de décongélation des poids. Par limite de mémoire GPU, les poids de l'encodeur ViT**l**14_reg n'ont pas pu être optimisés durant la tâche de segmentation.

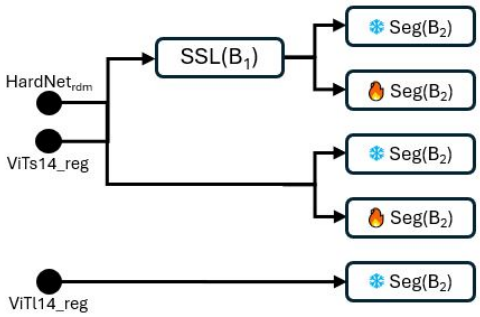

FIGURE 3 – Les scénarios d'entraînements explorés : le flocon signifie que l'encodeur reste figé tandis que la flamme désigne son optimisation via la décongélation des poids.

4.2 Implémentation

Les expériences décrites ont été réalisées à l'aide d'une carte graphique NVIDIA RTX A6000 de 48 Go de mémoire. Les entraînements SSL à la DINO sont réalisés à l'aide de la librairie lightly [5]. Une augmentation à la volée est effectuée durant les 300 époques des entraînements SSL avec un mini-batch de 128 images. Chacune aboutit à la création de 8 vues. La résolution des 2 vues "globales" est fixée à 224x224 et à 98x98 pour les 6 vues "locales". La tête de projection est composée de 3 couches linéaires. Sa dimension d'entrée dépend de la dimension du tenseur de sortie de chaque encodeur tandis que les dimensions des autres couches restent inchangées entre les expériences et valent respectivement 512, 128 et 2048.

Pour les entraînements supervisés sur B_2 suivant un préentraînement SSL sur B_1, les poids de départ des encodeurs correspondent à ceux ayant minimisé la fonction de coût sur les données de validation SSL. La segmentation est réalisée sur 150 époques en 5 folds cross validation avec le décodeur Lawin : 4 tenseurs caractéristiques sont extraits de l'encodeur et sont adaptés aux 4 entrées attendues du décodeur. La taille du mini-batch dépend de la taille mémoire GPU occupée par l'encodeur : 12 pour HardNet-MSEG et pour les encodeurs ViT lorsque leurs poids sont congelés mais 2 pour ViT**s**14_reg lors de sa décongélation progressive. Les performances sur l'ensemble test seront évaluées avec la métrique Dice, une mesure couramment utilisée en segmentation pour évaluer la similarité entre la prédiction du modèle et la vérité terrain. Elle se calcule par la formule suivante où *pred* désigne les pixels prédits par le modèle comme appartenant à la plaie et *vt* les pixels désignés par l'annotateur comme appartenant à une plaie :

$$DICE = \frac{|pred \cap vt|}{|pred| + |vt|}$$

4.3 Résultats

Durant les préentraînements SSL, un phénomène de surapprentissage apparaît au bout d'une centaine d'époques, quel que soit l'encodeur. La stratégie d'arrêt précoce est réglée à 30 époques pour limiter le temps de calcul. Les durées des préentraînements ont été d'environ 30 h pour une centaine d'époque. Quel que soit le scénario, l'optimisation des algorithmes utilisant HardNet varie entre 1 h et 2 h en fonction de la quantité de données. Le temps d'optimisation des scénarios avec ViT**s**14_reg varie entre 1 h et 3 h en fonction de la quantité de données mais aussi de l'état figé ou décongelé de l'encodeur. Ces temps sont similaires pour l'optimisation du scénario avec ViT**l**14_reg. Le tableau 1 présente les performances obtenues par les différents scénarios d'entraînements proposés en fonction de la métrique Dice sur la tâche de segmentation sur B_2.

4.4 Discussion

Dans le tableau 1, les lignes 6 et 8 montrent que l'augmentation de l'échelle du modèle ViT préentraînée "à la DINO" permet une amélioration des performances. Le passage de ViT**s**14_reg à ViT**l**14_reg est donc lié à une amélioration des capacités d'extraction des caractéristiques des images de plaies chroniques. Cependant, d'après les lignes 3 et 8, un modèle convolutif léger tel que HardNet-MSEG, optimisé sur B_1 et sans préentraînement SSL sur B_2 au préalable,

5. https ://github.com/lightly-ai/lightly

Encodeur	SSL(B_1)	Optimisation (B_2)	Train=25%	Train=50%	Train=70%	ligne
HardNet-MSEG$_{rdm}$	✔	❄	0.724±0.032	0.737±0.011	0.755±0.017	1
		🔥	0.756±0.034	0.784±0.009	0.798±0.013	2
	✘	🔥	0.694±0.038	0.736±0.033	0.771±0.023	3
ViTs14_reg	✔	❄	0.594±0.055	0.644±0.035	0.665±0.034	4
		🔥	0.674±0.017	0.709±0.023	0.720±0.025	5
	✘	❄	0.571±0.043	0.646±0.028	0.653±0.020	6
		🔥	0.685±0.025	0.721±0.021	0.729±0.006	7
ViTl14_reg	✘	❄	0.637±0.042	0.639±0.024	0.701±0.032	8

TABLE 1 – Performances en DICE des modèles sur la tâche de segmentation de B_2 : le flocon signifie que l'encodeur reste figé tandis que la flamme désigne son optimisation via la décongélation des poids.

prévaut sur ViTl14_reg initialisé via SSL DINO sur LVD-142M, quel que soit la quantité de données d'entraînement durant la tâche finale supervisée. Cela signifie que LVD-142M n'est pas adaptée pour des applications cliniques aussi spécifiques que l'analyse des plaies chroniques. Par ailleurs, quel que soit le choix de l'encodeur, de l'optimisation sur B_2 et de la quantité de données d'entraînement de B_2, un préentraînement via la méthodologie SSL DINO sur les images de plaies chroniques B_1 s'accompagne d'une amélioration de la métrique Dice en segmentation. Cette amélioration peut être source d'économies d'annotations. En effet, comme le montre les lignes 2 et 3, le modèle HardNet-MSEG préentraîné en SSL DINO sur B_1 puis optimisé en segmentation sur B_2 avec uniquement 25% de données annotées obtient des performances similaires avec le même encodeur mais sans préentraînement et ayant 70% de données annotées pour son entraînement en segmentation.

5 Conclusion

Dans cet article, nous avons évalué l'intérêt d'utiliser un encodeur entraîné "à la DINO" sur une base de données clinique spécifique puis sur une tâche de segmentation. Nous avons comparé dans différents scénarios d'entraînements les performances de l'encodeur ViT, dont les poids sont issus de l'article Dinov2, à l'encodeur léger HardNet-MSEG, dont les poids sont initialisés aléatoirement. Les résultats montrent qu'il n'est pas utile d'employer des architectures DINO préentraînées sur LVD-142M car des modèles légers peuvent être supérieurs sur des tâches spécifiques. De plus, le préentraînement d'un encodeur par du SSL "à la DINO" sur une base spécifique avec peu d'annotations présente un intérêt. Il serait intéressant de poursuivre ce travail en étudiant l'impact de l'augmentation de la base de données pour l'entraînement SSL grâce à l'ajout de toutes les bases de données publiques liées aux lésions dermatologiques.

Nous remercions l'ANRT ainsi que le réseau Cicat-Occitanie pour financer et soutenir la thèse CIFRE.

Références

[1] Mathilde CARON et al. "Emerging properties in self-supervised vision transformers". In : *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, p. 9650-9660.

[2] Ping CHAO et al. "Hardnet : A low memory traffic network". In : *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, p. 3552-3561.

[3] Alexey DOSOVITSKIY et al. "An image is worth 16x16 words : Transformers for image recognition at scale". In : *arXiv preprint arXiv :2010.11929* (2020).

[4] Gao HUANG et al. "Densely connected convolutional networks". In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, p. 4700-4708.

[5] Connah KENDRICK et al. "Translating clinical delineation of diabetic foot ulcers into machine interpretable segmentation". In : *arXiv preprint arXiv :2204.11618* (2022).

[6] Ting-Yu LIAO et al. "HarDNet-DFUS : Enhancing Backbone and Decoder of HarDNet-MSEG for Diabetic Foot Ulcer Image Segmentation". In : *Diabetic Foot Ulcers Grand Challenge*. Springer, 2022, p. 21-30.

[7] Maxime OQUAB et al. "Dinov2 : Learning robust visual features without supervision". In : *arXiv preprint arXiv :2304.07193* (2023).

[8] Utku OZBULAK et al. "Know Your Self-supervised Learning : A Survey on Image-based Generative and Discriminative Training". In : *arXiv preprint arXiv :2305.13689* (2023).

[9] Shaoqing REN et al. "Faster r-cnn : Towards real-time object detection with region proposal networks". In : *Advances in neural information processing systems* 28 (2015).

[10] Haotian YAN et al. "Lawin transformer : Improving semantic segmentation transformer with multi-scale representations via large window attention". In : *arXiv preprint arXiv :2201.01615* (2022).

[11] Chuyan ZHANG et al. "Dive into the details of self-supervised learning for medical image analysis". In : *Medical Image Analysis* 89 (2023), p. 102879.

Comparer le paradigme spatial au spatio-temporel pour estimer l'évolution d'indicateurs socio-économiques à partir d'images satellites

R. Jarry[1], M. Chaumont[1, 2], L. Berti-Équille[3], G. Subsol[1]
[1] LIRMM, Univ. Montpellier, CNRS, Montpellier, France
[2] Université de Nîmes, France
[3] ESPACE-DEV, Univ. Montpellier, IRD, UA, UG, UR, Montpellier, France

1 Introduction

Ces dernières années, beaucoup de travaux de recherche sur l'estimation de la pauvreté à partir d'images satellites ont été proposés (voir par exemple [1]). Les méthodes utilisant l'apprentissage profond donnent des résultats permettant d'estimer la pauvreté avec une précision correcte dans des pays où il est difficile de mener des enquêtes de terrain. Pour autant, des travaux récents observent des difficultés pour estimer *l'évolution de la pauvreté*, c'est-à-dire, estimer les variations de la pauvreté sur une période de temps donnée [2]. Les méthodes existantes ne reposent que sur des données spatiales, et ne sont pas adaptés à estimer des évolutions [3]. Une idée pour améliorer ces résultats serait de considérer la dépendance temporelle. Par exemple, dans [2], les auteurs notent une amélioration importante en estimant l'évolution de la pauvreté avec deux observations à deux dates différentes de la même zone géographique. Nous proposons d'étendre cette idée à des séries temporelles d'images satellites (SITS), qui sont des observations répétées d'une même zone géographique, à des dates variables. Cette idée s'appuie sur la réussite de l'utilisation des SITS dans d'autres domaines applicatifs [4]. Nous proposons d'évaluer si l'utilisation des SITS permet d'améliorer les estimations de l'évolution de la pauvreté.

2 Comparer un modèle spatial à un modèle spatio-temporel

Deux zones d'études : Zanzibar et Damas. Les zones d'études choisies sont exposées 2. La zone d'étude (1) est un voisinage large autour de Zanzibar, en Tanzanie, contenant, entre autres, Dar Es Salam, la capitale économique du pays. Cette zone est d'intérêt applicatif dans le cadre de ces recherches. La zone d'étude (2) est un voisinage de la ville de Damas en Syrie, avec une partie du Liban, notamment la ville de Beyrouth. Elle est sélectionnée, car elle contient à la fois des zones de guerres, sur lesquelles on peut observer des diminutions de l'intensité lumineuse nocturne (ILN), avec des zones qui ont suivi un développement économique normal. Nous étudierons ces zones de 2000 à 2020.

SITS & ILN. Nous avons choisi d'utiliser les images satellites multispectrales Landsat-7, parce qu'elles couvrent l'ensemble du globe avec une résolution spatiale et temporelle acceptable (30 mètres et 16 jours respectivement). Nous construisons à partir de ces observations des composites annuels, c'est-à-dire des observations moyennes (sans nuages) de notre zone d'étude pour chaque année. Nous avons choisi d'utiliser les intensités lumineuses nocturnes (ILN) de [5] comme données de référence à estimer, pour les mêmes raisons de couverture spatio-temporelle de Landsat-7, mais aussi, car les ILNs sont un proxy standard pour certains indicateurs socio-économiques [6]. Pour que la taille des images d'ILN de notre zone d'étude coïncide avec la taille des images Landsat-7, les ILNs sont sur-échantillonnées avec la méthode du plus proche voisin.

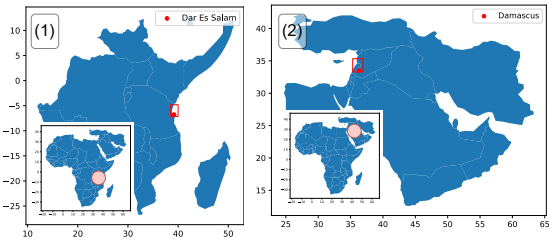

FIGURE 1 – La zone (1), à gauche, est un voisinage de Zanzibar. La zone (2), à droite, est un voisinage de Damas.

La base d'apprentissage. Le processus de construction du jeu de données d'apprentissage est décrit en figure 2. Les SITS et séries d'ILN de la zone d'étude sont découpées en

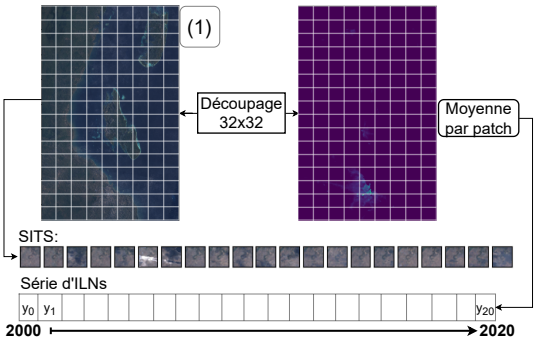

FIGURE 2 – Collecte des données et constitution du jeu d'apprentissage sur la zone (1). Le processus est identique sur la zone (2). Notons que les mailles de la grille sont plus petites en réalité.

Score		$MAE \downarrow (1)$	$R^2 \uparrow (1)$	$MAE \downarrow (2)$	$R^2 \uparrow (2)$
Par année	ST	$\mathbf{0.085}^{\pm \mathbf{0.010}}$	$\mathbf{0.695}^{\pm \mathbf{0.063}}$	$\mathbf{1.011}^{\pm \mathbf{0.061}}$	$\mathbf{0.543}^{\pm \mathbf{0.014}}$
	S	$0.098^{\pm 0.008}$	$0.591^{\pm 0.040}$	$1.262^{\pm 0.104}$	$0.431^{\pm 0.028}$
$\Delta t = 1$	ST	$\mathbf{0.033}^{\pm \mathbf{0.002}}$	$\mathbf{0.123}^{\pm \mathbf{0.062}}$	$\mathbf{0.453}^{\pm \mathbf{0.020}}$	$\mathbf{0.128}^{\pm \mathbf{0.010}}$
	S	$0.087^{\pm 0.009}$	$-4.664^{\pm 1.390}$	$0.685^{\pm 0.066}$	$-1.240^{\pm 0.439}$
$\Delta t = 10$	ST	$\mathbf{0.103}^{\pm \mathbf{0.005}}$	$\mathbf{0.322}^{\pm \mathbf{0.042}}$	$\mathbf{1.083}^{\pm \mathbf{0.043}}$	$\mathbf{0.218}^{\pm \mathbf{0.013}}$
	S	$0.140^{\pm 0.010}$	$-0.291^{\pm 0.083}$	$1.287^{\pm 0.056}$	$-0.150^{\pm 0.099}$
$\Delta t = 15$	ST	$\mathbf{0.137}^{\pm \mathbf{0.009}}$	$\mathbf{0.439}^{\pm \mathbf{0.048}}$	$\mathbf{1.069}^{\pm \mathbf{0.041}}$	$\mathbf{0.278}^{\pm \mathbf{0.020}}$
	S	$0.163^{\pm 0.010}$	$-0.032^{\pm 0.129}$	$1.314^{\pm 0.059}$	$-0.123^{\pm 0.071}$

TABLE 1 – Score MAE et R^2 sur les zones (1) et (2). Les exposants sont les écarts-types sur les cinq expériences.

patchs, selon une grille régulière dont chaque maille fait 32×32 pixels. Les patchs d'ILN de chaque pas de temps sont moyennés, résultant en une série de valeurs d'ILN sur la période temporelle d'étude. Un exemple d'apprentissage est alors une paire constituée d'une série de patchs d'image Landsat-7 d'une même maille du quadrillage et l'évolution de l'ILN sur cette maille. Nos expériences sont menées avec une validation croisée à cinq plis. Il y a environ 40 000 exemples d'apprentissage dans chacune des zones (1) et (2).

Les architectures *Transformer*. Nous avons décidé de travailler avec l'architecture Transformer, car c'est une architecture de l'état de l'art ayant été adaptée à la fois pour des images, mais aussi des séquences d'images. Deux modèles sont construits, sur la base des travaux de [7] pour le modèle spatial (S), et des travaux de [4] pour le modèle spatio-temporel (ST). Le modèle spatial traite chaque image d'une séquence indépendamment des autres et estime une unique ILN par image. Nous l'utiliserons séquentiellement sur toutes les images d'une même SITS pour obtenir une prédiction d'évolution. Le modèle spatio-temporel analyse toutes les images de la SITS en une seule fois, recherchant des motifs spatiaux et temporels à corréler avec l'évolution à estimer.

3 Résultats

Par années. Dans la première ligne de score du tableau 1, nous calculons la moyenne des scores pour chacune des années de 2000 à 2020. Dans ce contexte, nous remarquons tout d'abord que les deux modèles atteignent des scores assez équivalents lorsqu'ils prédisent des valeurs d'ILN pour une année donnée, soit $R^2 = 0.69$ pour le modèle ST et $R^2 = 0.59$ pour le modèle S. Le score MAE est bien plus élevé sur la zone (2) que sur la zone (1), alors que les scores R^2 sont du même ordre (même si légèrement inférieurs sur la zone (2)), suggérant qu'il est plus difficile d'estimer l'ILN sur la zone (2) que sur la zone (1).

Par évolutions. Dans le reste du tableau 1, nous calculons la moyenne des scores sur les *évolutions* espacées de Δt années. Pour chaque exemple du JDD, l'évolution de l'ILN est obtenu en calculant la différence d'ILN à l'année $t + \Delta t$ et t. Nous observons que pour $\Delta t = 1$, *i.e.* une évolution d'un an, tous les scores R^2 sont soit négatifs, soit proche de 0 ce qui signifie que les dépendances à court terme sont difficiles à capturer dans les domaines spatial et temporel. Nous pensons que cela est dû au fait que les variations annuelles de l'ILN sont difficilement perceptibles dans les données. Cependant, à mesure que Δt augmente, le modèle ST prédit mieux l'évolution avec $R^2 = 0,44$ pour $\Delta t = 15$ sur la zone (1). D'autre part, le modèle S reste peu fiable puisque $R^2 < 0$ pour tous les Δt. Individuellement, les scores R^2 des deux modèles s'améliorent lorsque Δt augmente. En revanche, ce dernier point n'est pas vérifié pour le score MAE, qui augmente dès que Δt augmente. En effet, les évolutions de l'ILN d'une année à l'autre n'ont pas la même amplitude que les évolutions de l'ILN sur une période de 10 ou 15 ans. Par conséquent, il est impossible de comparer les scores MAE pour des Δt différents.

4 Perspectives

Nous obtenons des résultats encourageants, qui nous permettent d'observer une supériorité du modèle spatio-temporel sur le modèle spatial. Néanmoins, nous avons fait notre analyse sur deux zones d'étude restreintes. De plus, les scores obtenus (notamment R^2) sont encore bien trop faibles pour une utilisation pratique de tels modèles. Une perspective est d'agrandir cette zone d'étude, à un continent entier par exemple. Cela permettrait de rendre notre analyse plus robuste et d'améliorer les performances des modèles.

Informations - Cet article est un résumé en français, avec des expériences complémentaires, de l'article suivant : R. Jarry, M. Chaumont, L. Berti-Équille, G. Subsol. "Comparing Spatial and Spatio-Temporal Paradigms to Estimate The Evolution of Socio-Economical Indicators from Satellite Images." in *IGARSS 2023 - IEEE International Geoscience and Remote Sensing Symposium* (p. 5790-5793). 10.1109/IGARSS52108.2023.10282306

Références

[1] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, August 2016.

[2] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," *Nature Communications*, vol. 11, no. 1, pp. 2583, May 2020.

[3] L. Kondmann and X. X. Zhu, "Measuring changes in poverty with deep learning and satellite images," *ICLR, Practical ML for Developing Countries*, p. 6, 2020.

[4] Y. Yuan, L. Lin, Q. Liu, R. Hang, and Z. Zhou, "SITS-Former : A pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification," *Intl Journal of Applied Earth Observation and Geoinformation*, vol. 106, pp. 102651, Feb. 2022.

[5] Z. Chen, B. Yu, C. Yang, Y. Zhou, S. Yao, X. Qian, C. Wang, B. Wu, and J. Wu, "An extended time series (2000–2018) of global NPP-VIIRS-like nighttime light data from a cross-sensor calibration," *Earth System Science Data*, vol. 13, no. 3, pp. 889–906, March 2021.

[6] A. M. Noor, V. A. Alegana, P. W. Gething, A. J. Tatem, and R. W. Snow, "Using remotely sensed night-time light as a proxy for poverty in Africa," *Population Health Metrics*, vol. 6, no. 1, pp. 5, 2008.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale," in *ICLR*, 2021.

Session 2 : Planification et logique

Vers une approche floue pour le design de plan expérimental

O. Rousselle, J.-P. Poli, N. Ben Abdallah

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

{prenom.nom}@cea.fr

Résumé

Nous présentons dans cet article une approche floue interprétable du design expérimental sous contraintes qui peut être utilisée avec peu de données. L'objectif est de fournir aux expérimentateurs un algorithme leur permettant d'échantillonner de manière optimale. Nous détaillons les différentes étapes de notre algorithme qui consiste à recommander la prochaine expérience à réaliser et à construire une base de règles floues de Sugeno. Nous présentons ensuite quelques résultats de notre algorithme, que nous comparons avec l'approche bayésienne.

Mots-clés

Apprentissage actif, design expérimental, flou, règles, optimisation, interprétabilité.

Abstract

In this paper we present an interpretable fuzzy approach to constrained experimental design that can be used with little data. The objective is to provide experimenters with an algorithm allowing them to sample optimally. We detail the different steps of our algorithm which consists of recommending the next experiment to be carried out and constructing a Sugeno fuzzy rule base. We will then present some results of our algorithm, which we compare with the Bayesian approach.

Keywords

Active learning, experimental design, fuzzy, rules, optimization, interpretability.

1 Introduction

Les sciences expérimentales nécessitent d'explorer un espace de possibilités souvent très vaste pour s'approcher d'un optimum. Classiquement, l'approche par essais et erreurs est utilisée dans ce domaine pour collecter des données expérimentales, dont la production peut être coûteuse et longue. Le processus est répété jusqu'à ce qu'une propriété ou une performance désirée soit atteinte.

Différentes méthodes d'échantillonnage sont utilisées dans la recherche expérimentale, telles que l'échantillonnage aléatoire, l'échantillonnage factoriel, la méthode des surfaces de réponse, l'optimisation bayésienne [1], l'algorithme de couverture optimale [2], etc. Les réseaux de neurones ont également été utilisés pour la recherche expérimentale [3], mais ont l'inconvénient d'être une boîte noire.

Sans perte de généralité, nous prenons comme exemple la découverte des matériaux, qui vise à produire des matériaux performants pour un usage ciblé. Ces matériaux sont généralement produits à partir d'un mélange de composés initiaux soumis à un certain procédé de fabrication. Ces dernières années, l'intelligence artificielle a accéléré l'innovation dans ces domaines [4, 5, 6].

Dans ce contexte, l'objectif de nos travaux est de développer une méthode basée sur la logique floue appliquée au plan expérimental. Nous définissons notre problème comme tester différents ensembles de paramètres (c'est-à-dire la composition d'un matériau et les paramètres du procédé) pour maximiser une propriété donnée (par exemple la robustesse de ce matériau). En particulier, nous souhaitons trouver une méthode automatique pour échantillonner de manière itérative et optimale les paramètres expérimentaux.

Notre motivation est de fournir un outil pour aider les expérimentateurs à déterminer quels sont les prochains ensembles de paramètres à tester et qui fonctionne avec peu de données. Nous visons à réduire le nombre d'expérimentations pour atteindre une performance cible, à la fois pour réduire le gaspillage de matières premières et converger plus rapidement vers un matériau innovant. Pour garder l'expert humain dans la boucle, nous prêtons attention à l'interprétabilité, en donnant des indices à l'utilisateur sur le choix de la prochaine configuration expérimentale. En effet, l'explicabilité/interprétabilité vise à rendre le fonctionnement et les résultats du modèle plus intelligibles et transparents pour les humains, afin de renforcer la confiance dans la prise de décision et ainsi son acceptabilité.

Le document est structuré comme suit. La section suivante donne un aperçu de l'approche. La section 3 décrit la méthode de régression qui se rapproche de la fonction objectif. La section 4 explique le processus derrière la sélection de la prochaine expérience à réaliser. Nous montrons les résultats et la comparaison avec l'optimisation bayésienne dans la Section 6. Notre approche étant dédiée aux experts humains, la Section 7 présente la manière dont l'utilisateur final est considéré. Enfin, nous tirons quelques conclusions et perspectives.

2 Vue d'ensemble de l'approche

Pour satisfaire les besoins des expérimentateurs, nous avons conçu une approche basée les principes suivants :

— Elle doit mettre en œuvre un échantillonnage adap-

tatif [7], c'est-à-dire un processus séquentiel qui décide de l'emplacement du point suivant en équilibrant les critères d'exploration et d'exploitation ;
— Elle doit être capable de combiner apprentissage (à partir de quelques données expérimentales) et modélisation de connaissances expertes ;
— Robustesse : de petits changements dans les points initiaux ne doivent pas entraîner de changements importants dans les résultats et les prédictions ;
— Interprétabilité : les étapes et les résultats du modèle doivent être intelligibles pour les humains, pour renforcer la confiance dans la prise de décision ;
— Notre approche doit pouvoir travailler sur des problèmes de mélanges de grande dimension.

Pour répondre à ces prérequis, nous avons développé une approche dont les différentes étapes sont détaillées dans la Fig. 1.

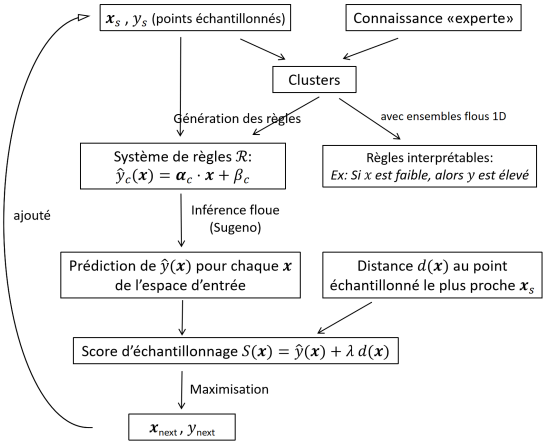

FIGURE 1 – Différentes étapes de l'approche proposée.

Le point expérimental sélectionné est celui qui maximise le score d'échantillonnage (section 4). Ce processus est répété sur plusieurs itérations, chaque itération correspondant à une expérience.

Dans ce travail, nous considérons que les propriétés des matériaux sont données sous forme de valeurs réelles (souvent bornées). Nous devons donc construire une base de règles floues pour une régression (voir section 3). Nous avons choisi d'utiliser un système d'inférence flou de type Sugeno pour son efficacité et pour le fait qu'il fonctionne avec les techniques d'optimisation adaptatives.

Avec ce choix, nous avons privilégié les performances de la prédiction plutôt que l'interprétabilité. Pour compenser, nous proposons une méthode pour extraire un substitut plus interprétable du modèle (section 5).

Remarque : les notations utilisées dans cet article sont détaillées en annexe.

3 Algorithme de régression basé sur le clustering flou

Considérons le problème de la prédiction de la propriété d'un matériau, $\hat{y}$, pour chaque point de l'espace d'entrée (c'est-à-dire tout mélange de matières premières et de paramètres de processus). Nous avons basé notre approche sur plusieurs travaux antérieurs [8, 9, 10] qui partagent l'utilisation d'une méthode de clustering comme première étape dans l'induction de règles. Notre méthode diffère légèrement dans le sens où elle est destinée à des problèmes éventuellement de grande dimension. De plus, nous avons amélioré la manière dont les fonctions d'appartenance sont apprises et le calcul des coefficients de régression pour les conclusions de la règle de Sugeno.

3.1 Clustering

Tout d'abord, nous effectuons un clustering flou multidimensionnel des points déjà échantillonnés $\boldsymbol{x}_s$ pour mettre en évidence différents groupes. Rappelons que la particularité du clustering flou est qu'un point peut appartenir à plusieurs clusters, avec éventuellement des degrés d'appartenance différents.

Le clustering flou s'applique à la fois aux variables d'entrée et de sortie, et chaque cluster c comprend un centre noté $\boldsymbol{m}_c$. Nous notons C l'ensemble des clusters. Le nombre de clusters n_c est souvent un hyperparamètre, dont la valeur doit être définie en respectant les considérations suivantes :

— Trop de clusters peuvent conduire à un surapprentissage. Le modèle s'est bien adapté aux données d'entraînement (points déjà échantillonnés), mais il peut avoir du mal à se généraliser à de nouvelles données. De plus, avoir trop de clusters implique trop de paramètres, et l'optimisation décrite plus loin dans cet article ne sera pas possible.
— Un faible nombre de clusters peut faciliter l'interprétation du modèle et réduire la complexité des calculs (moins de paramètres à optimiser).

Le nombre de clusters peut rester constant au cours des différentes itérations de l'expérience ou peut augmenter régulièrement par paliers en fonction du nombre de points déjà échantillonnés.

Nous nous intéressons maintenant à mesurer le degré d'appartenance d'un point donné $\boldsymbol{x}$ dans l'espace d'entrée à chaque cluster, noté $\mu_c(\boldsymbol{x})$. Nous avons choisi de construire une fonction d'appartenance qui dépend de plusieurs variables d'entrée. L'avantage est de prendre en compte l'interaction entre les variables et d'obtenir une partition forte multidimensionnelle :

$$\forall \boldsymbol{x}\ ,\ \sum_{c \in C} \mu_c(\boldsymbol{x}) = 1. \tag{1}$$

Les degrés d'appartenance peuvent être calculés en s'inspirant de l'algorithme FCM (Fuzzy Clustering Means) [11], c'est-à-dire en minimisant la fonction objectif [12] :

$$\sum_{\boldsymbol{x}_s} \sum_{\boldsymbol{m}_c} \mu_c(\boldsymbol{x}_s) ||\boldsymbol{x}_s - \boldsymbol{m}_c||^2 \tag{2}$$

avec $\boldsymbol{x}_s$ les points considérés, m_c les centres de chaque cluster, et $\mu_c(\boldsymbol{x}_s)$ le coefficient d'appartenance du point $\boldsymbol{x}_s$ au cluster c. Les degrés d'appartenance obtenus sont :

$$\mu_c(\boldsymbol{x}_s) = \sum_{c'} \left(\frac{||\boldsymbol{x}_s - \boldsymbol{m}_c||^2}{||\boldsymbol{x}_s - \boldsymbol{m}_{c'}||^2} \right)^{-\frac{2}{m-1}} \tag{3}$$

avec m le paramètre "fuzzifier" qui influence le flou de la partition. Il va de 1 (partition nette) à $+\infty$. Généralement, m est choisi égal à 2. Chaque point $\boldsymbol{x}$ se voit ensuite attribuer un degré d'appartenance à chaque cluster. Des exemples de degrés d'appartenance sont illustrés dans le diagramme ternaire sur la Fig. 2 avec 3 clusters et avec 3 variables d'entrée x_1, x_2, x_3. En guise de lecture du diagramme ternaire, la croix rouge indique par exemple le point $(x_1, x_2, x_3) = (0.55, 0.2, 0.25)$.

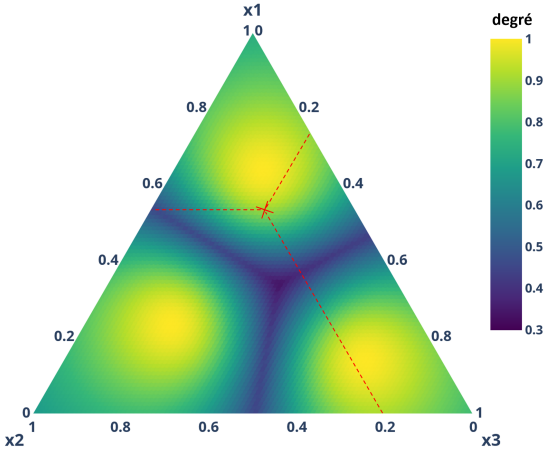

FIGURE 2 – Degrés d'appartenance - étude de cas avec 3 variables de composition et 3 clusters.

La méthode ne vise pas à caractériser les clusters, mais à les décrire dans leur ensemble. La distribution associée correspond donc à des degrés d'appartenance qui indiquent dans quelle mesure un point appartient à chaque cluster [13].
Remarque : on peut également construire une fonction d'appartenance qui dépend uniquement de la distance au centre du cluster. Il peut prendre une forme triangulaire, gaussienne [8, 14] ou Cauchy [10]. Cependant, nous n'avons pas choisi une telle fonction d'appartenance car elle ne considère que les similarités internes au sein d'un cluster (et non les dissemblances externes avec d'autres clusters), et parce que la partition floue n'est pas forte (la somme des degrés d'appartenance n'est pas égale à 1).

3.2 Génération de règles floues

Chaque cluster/région c est à l'origine d'une règle R. Chaque règle R est caractérisée par sa région antécédente multidimensionnelle et ses coefficients de régression $(\boldsymbol{\alpha}_c, \beta_c)$, sous la forme :

$$\hat{y}_c(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_{c_i} x_i + \beta_c = \boldsymbol{\alpha}_c \cdot \boldsymbol{x} + \beta_c. \tag{4}$$

Les coefficients $(\boldsymbol{\alpha}_c, \beta_c)$ de chaque règle sont déterminés par le processus d'optimisation décrit plus loin dans cet article.

Règle experte Une règle peut également provenir d'une expertise humaine ou être issue de la littérature. Par exemple, un expert peut indiquer une autre région d'intérêt en ajoutant un autre centre $\boldsymbol{m}_c$. Ce cluster générera également une règle caractérisée par sa région antécédente et ses coefficients de régression $(\boldsymbol{\alpha}_c, \beta_c)$. Cette règle sera ensuite ajoutée au système de règles déjà généré à partir des données.

Cas des variables discrètes Dans le cas de variables discrètes ou catégorielles, il est nécessaire de transformer ces données pour pouvoir intégrer ces variables dans la régression. Pour ce faire, nous utilisons l'encodage One-Hot, qui consiste à transformer la variable en plusieurs variables binaires, où chaque variable binaire représente une catégorie unique de la variable d'origine.

Processus récursif pour le partitionnement flou A chaque nouveau point échantillonné, nous évaluons à quel cluster il appartient, c'est-à-dire le cluster ayant le plus haut degré d'appartenance. Le centre $\boldsymbol{m}_c$ du cluster c est ensuite modifié en calculant la moyenne pondérée suivante (en notant $\boldsymbol{x}_{last}$ le dernier point testé) [15] :

$$\boldsymbol{m}'_c = \frac{\boldsymbol{m}_c + \mu_c(\boldsymbol{x}_{last})\, \boldsymbol{x}_{last}}{\mu_c(\boldsymbol{x}_{last})}. \tag{5}$$

3.3 Optimisation des coefficients de régression

Le résultat de notre modèle est généré en combinant les fonctions linéaires des règles, comme dans les systèmes de Sugeno :

$$\hat{y}(\boldsymbol{x}) = \sum_{c} \mu_c(\boldsymbol{x})\, \hat{y}_c(\boldsymbol{x}) \tag{6}$$

avec $\hat{y}_c(\boldsymbol{x}) = \boldsymbol{\alpha}_c \cdot \boldsymbol{x} + \beta_c$.
Nous déterminons les coefficients de régression optimaux $(\boldsymbol{\alpha}_c, \beta_c)$ en minimisant l'écart au carré entre les valeurs prédites des points déjà échantillonnés $\hat{y}$ et leurs valeurs réelles y . Chaque cluster contient $N + 1$ coefficients à optimiser. Au total, il y a donc $n_c(N + 1)$ paramètres à optimiser, il nous faut donc au moins $n_c(N + 1)$ points pour réaliser cette optimisation.
Pour chaque point déjà échantillonné, nous prédisons la valeur de sortie à l'aide de la formule d'inférence, que nous comparons à la valeur de sortie réelle. L'objectif est de minimiser la différence entre les valeurs prédites $\hat{y}$ et les valeurs réelles y. On cherche donc à minimiser la fonction de perte :

$$\begin{aligned} L(\boldsymbol{\alpha}, \beta) &= \sum_{\boldsymbol{x}_s} \left(\hat{y}(\boldsymbol{x}_s) - y(\boldsymbol{x}_s)\right)^2 \\ &= \sum_{\boldsymbol{x}_s} \left(\sum_{c} \mu_c(\boldsymbol{x}_s)\, (\boldsymbol{\alpha}_c \cdot \boldsymbol{x}_s + \beta_c) - y(\boldsymbol{x}_s) \right)^2 \end{aligned} \tag{7}$$

3.4 Prédiction de la valeur de sortie pour chaque point d'entrée

Pour chaque point d'entrée $\boldsymbol{x}$, nous prédisons la valeur de sortie $\hat{y}$. La figure 3 montre un exemple de valeurs prédites $\hat{y}$ obtenues avec notre algorithme pour un cas avec 3 variables d'entrée.

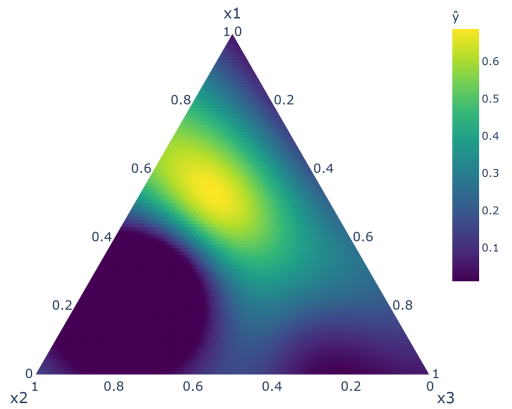

FIGURE 3 – Diagramme ternaire illustrant un exemple de valeurs prédites $\hat{y}$ pour un cas avec 3 variables d'entrée.

Nous pouvons évaluer la précision du modèle d'inférence en utilisant les points déjà échantillonnés. Pour chaque point échantillonné, nous prédisons la valeur de sortie $\hat{y}$ à l'aide du système d'inférence, que nous comparons à la valeur réelle y. On peut alors calculer l'écart au carré moyen RMSE et le coefficient de détermination R^2 pour juger de la qualité du modèle. L'objectif est d'avoir un modèle qui maximise R^2 et minimise le RMSE. Les résultats sont présentés dans la partie 6.
Dans [10], les auteurs ont montré que l'algorithme de régression basé sur le clustering flou avec descente de gradient est plus efficace que les réseaux de neurones pour un problème de régression (prédiction de la forme d'une fonction) pour le cas 1D, et avec beaucoup moins d'hyperparamètres.

4 Sélection de la prochaine expérience

La méthode proposée choisit également le prochain point à échantillonner de manière déterministe, comme un compromis entre exploitation et exploration. L'exploitation signifie privilégier les régions à fort potentiel, c'est-à-dire avec des valeurs de sortie élevées prédites $\hat{y}$, tandis que l'exploration signifie explorer les régions encore non testées.
Nous devons introduire une variable qui capture cette partie exploration. Une variable naturelle pour cela est la distance euclidienne au point échantillonné le plus proche, notée d. Le calcul des distances entre chaque point de l'espace d'entrée et le point déjà échantillonné le plus proche est effectué de manière optimisée à l'aide de l'algorithme KD-tree [16]. Ensuite, pour chaque point d'entrée $\boldsymbol{x}$, $\hat{y}$ et d sont calculés ; $\hat{y}$ code l'exploitation, et d code l'exploration. Pour trouver un compromis entre exploitation et exploration, nous introduisons une nouvelle variable appelée "score d'échantillonnage" et notée S :

$$S(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) + \lambda\, d(\boldsymbol{x}) \tag{8}$$

où $\hat{y}$ et d sont ici normalisés, et où λ est un hyperparamètre.

λ peut être choisi constant ou il peut changer en fonction du nombre d'expériences réalisées. Augmenter λ favorisera l'exploration par rapport à l'exploitation.
Une illustration de S est présentée dans la Fig. 4 pour un exemple avec 3 variables d'entrée. Les zones en bleu sont celles autour des points déjà échantillonnés, et les zones en jaune sont les régions d'intérêt. Une structure en réseau peut être observée, due au compromis entre exploitation et exploration.
Le prochain point proposé par l'algorithme sera celui qui maximise $S(\boldsymbol{x})$. Par exemple, le prochain point à tester pourrait être

$$\{x_1 = 0.35\ ,\ x_2 = 0.5\ ,\ x_3 = 0.15\}\,. \tag{9}$$

Alternativement, l'algorithme peut également proposer une région d'intérêt à l'expérimentateur. Cette région comprend tous les points ayant un score d'échantillonnage S supérieur à un seuil donné (par exemple 0.9). Par exemple, une région d'intérêt pourrait être :

$$\begin{cases} 0.3 \leq x_1 \leq 0.425 \\ 0.46 \leq x_2 \leq 0.535 \\ 0.11 \leq x_3 \leq 0.21. \end{cases} \tag{10}$$

puis l'expérimentateur choisira un point dans cette région.
Une explication peut également être donnée pour justifier le point/région proposé. Par exemple : "l'algorithme propose ce point/région à explorer à côté d'une zone déjà exploitée et qui a donné de bons résultats."
La proportion exploitation/exploration du prochain point testé peut également être fournie à l'expérimentateur :

$$p_{\text{exploitation}} = \frac{\hat{y}(\boldsymbol{x}_{next})}{\hat{y}(\boldsymbol{x}_{next}) + \lambda\, d(\boldsymbol{x}_{next})} \tag{11}$$

$$p_{\text{exploration}} = 1 - p_{\text{exploitation}}. \tag{12}$$

Enfin, des contraintes peuvent être prises en compte pour restreindre l'espace d'entrée ; par exemple $x_2 < 0.5$. Expérimentalement, ces contraintes pourraient être imposées par le dispositif expérimental, par exemple du fait des limites de la machine expérimentale ou du fait de lois physico-chimiques (comme la loi de miscibilité) dans le cas d'un mélange de matériaux.

5 Interprétabilité

Les avantages d'avoir des règles multidimensionnelles par rapport aux règles basées sur des ensembles flous 1D sont la possibilité de contrôler le nombre de règles souhaité (correspondant au nombre de clusters), d'éviter l'explosion du nombre de règles par rapport au nombre de dimensions, et de capturer interactions entre variables. Cependant, les règles multidimensionnelles sont moins interprétables que le cas unidimensionnel. Nous pouvons rendre le système multidimensionnel plus interprétable en établissant différentes catégories linguistiques par variable (par exemple faible/moyen/élevé) et en projetant les centres de chaque cluster sur chaque axe de variable [14].

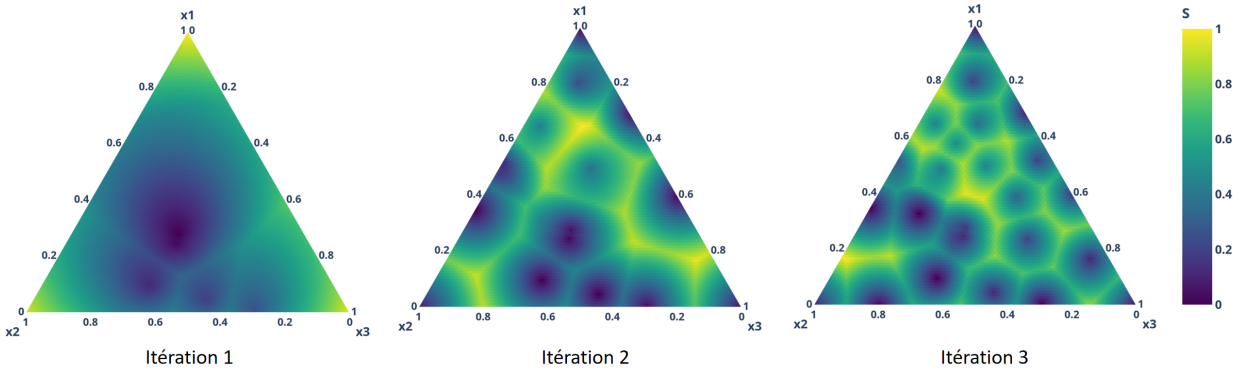

FIGURE 4 – Diagrammes ternaires illustrant l'évolution du score d'échantillonnage S à différentes itérations, pour un cas avec 3 variables d'entrée.

Notre algorithme d'interprétabilité suit les étapes suivantes :

— Nous divisons chaque variable d'entrée et de sortie en différents ensembles flous triangulaires/trapézoïdaux f, dont les sommets correspondent aux centres des clusters projetés sur chaque axe. Si deux ensembles flous sont trop proches (par exemple distance $<$ distance seuil), nous les fusionnons ;

— Pour une variable x_i $(i \in [1; N+1])$, en notant m_{c_i} la i^{ieme} composante du centre $\boldsymbol{m}_c$, alors le sous-ensemble associé au cluster R est celui avec le degré d'appartenance le plus élevé $\mu^f(m_{c_i})$.

Un cluster sera associé à $N+1$ sous-ensembles flous (un par variable).

La figure 5 illustre le système de règles interprétables obtenu pour un cas avec 1 variable d'entrée x et 3 clusters. Les règles 1D sont utilisées comme substituts des règles multidimensionnelles pour aider l'utilisateur à comprendre ce modèle.

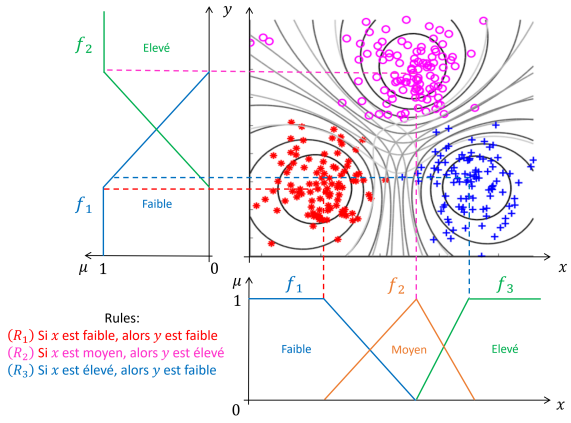

FIGURE 5 – Illustration de 3 clusters avec des degrés d'appartenance représentés par des lignes de niveau noires (tirées de [13]) et des ensembles flous f_i des variables x et y. L'association de chaque centre de cluster à ces ensembles flous 1D rend les règles plus interprétables.

6 Résultats expérimentaux

Dans cette section, nous présentons les résultats de différents tests pour caractériser notre algorithme. Par souci de reproductibilité, nous donnons d'abord quelques détails sur la mise en œuvre.

6.1 Considérations d'implémentation

Pour effectuer le clustering multidimensionnel, nous avons hybridé deux approches : nous utilisons la méthode de clustering hiérarchique [17] pour obtenir les centres des clusters et nous utilisons la fin de l'approche c-means floue [11] pour déterminer les fonctions d'appartenance. En effet, le clustering hiérarchique a l'avantage d'être totalement déterministe. Nous avons déterminé empiriquement le nombre de clusters pour chaque ensemble de données.

L'optimisation de la fonction de coût Eq. 7 peut être implémentée par la méthode Trust Region Reflective (TRF) [18], l'algorithme de Levenberg-Marquardt [9], l'algorithme des moindres carrés récursifs [8] ou descente de gradient [19, 10]. Nous avons utilisé l'algorithme TRF car il s'agit d'une méthode robuste (peu sensible au choix du point de départ), bien adaptée aux problèmes complexes avec des résidus non linéaires, adaptée aux grands problèmes clairsemés avec des bornes, et qui ne nécessite pas d'hyperparamètres supplémentaires.

6.2 Jeux de données jouet

Nous avons d'abord évalué notre méthode sur un jeu de données jouet que nous avons généré à partir d'une fonction sinus avec éventuellement plusieurs entrées. Un petit nombre d'entrées nous aide à visualiser les résultats pour les qualifier, tandis qu'un grand nombre permet de valider notre algorithme.

Fonction sinus à 1 variable d'entrée Nous avons d'abord testé notre algorithme de régression décrit dans la partie 3 avec le cas simple de la fonction $f(x) = \sin(2\pi x)$, avec 6 points initiaux et 2 clusters flous. Le résultat est tracé sur la Fig. 6, avec les valeurs de sortie prédites et les valeurs de sortie réelles.

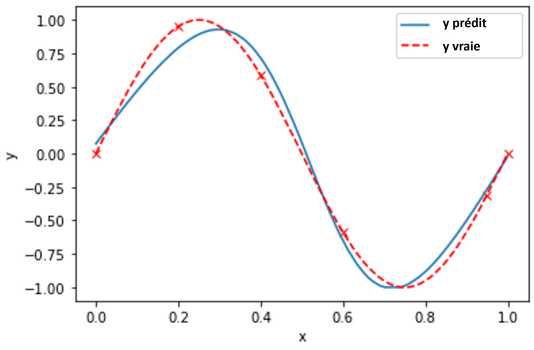

FIGURE 6 – Illustration de la sortie prédite $\hat{y}$ à l'aide de l'algorithme de régression basé sur le clustering flou.

Fonction sinus à 3 variables d'entrée Nous avons ensuite testé notre algorithme pour la fonction objectif sinus suivante avec 3 variables d'entrée :

$$f(\boldsymbol{x}) = \left| \prod_{k=1}^{3} \sin(k\pi x_k) \right|. \tag{13}$$

Cette fonction objectif a été choisie car elle présente une forme non triviale avec plusieurs maxima et un maximum global, avec des valeurs de sortie comprises entre 0 et 1, et parce qu'elle peut être tracée dans un diagramme ternaire (voir Fig. 7). Cela nous aide également à caractériser l'approche puisque nous n'aurons jamais un plan expérimental complet à partir d'une application réelle.

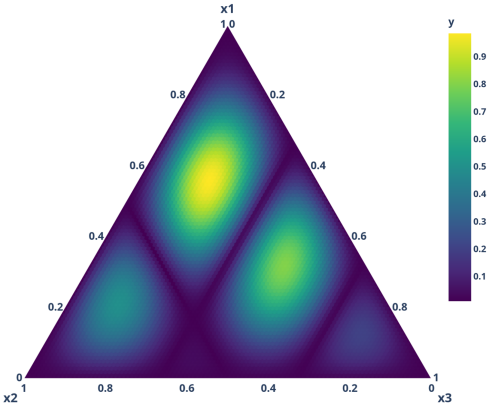

FIGURE 7 – Diagramme ternaire de la fonction objectif f donnée par l'équation (13).

Chaque axe x_i contient 100 valeurs de 0 à 1, donc au total nous avons 5151 points dans l'espace d'entrée. Notre objectif est de converger le plus rapidement possible vers la valeur optimale.
On choisit initialement 5 points aléatoires et on effectue 50 itérations (avec un point testé par itération). L'efficacité de notre algorithme est mesurée avec le critère du nombre d'itérations M nécessaires pour atteindre 80% de la valeur optimale de y (qui est de 0,984 dans notre cas d'étude). La figure 8 montre la meilleure valeur y obtenue parmi les points testés jusqu'à une itération donnée ; 80% de la valeur optimale est atteinte à l'itération $M = 24$.

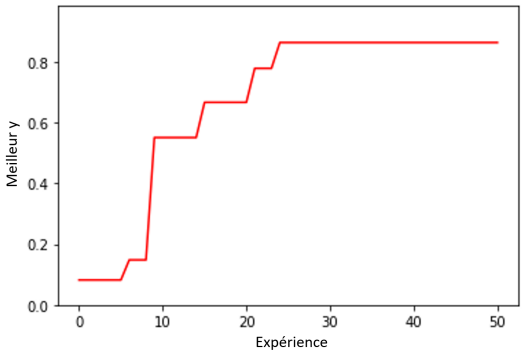

FIGURE 8 – Meilleure valeur de y obtenue pour chaque itération.

La proportion d'exploitation/exploration du point testé à chaque itération est tracée dans l'histogramme Fig. 9.

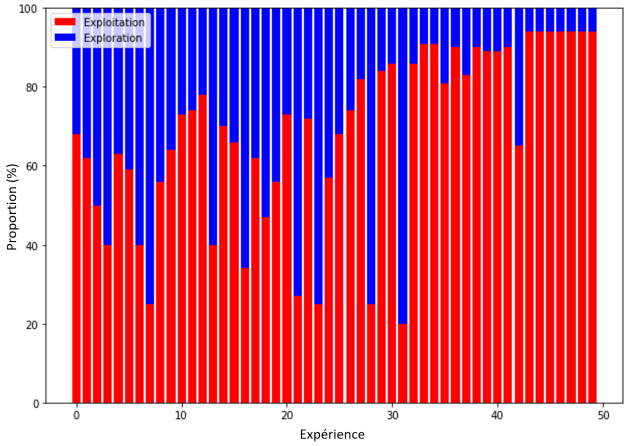

FIGURE 9 – Proportion exploitation/exploration pour chaque point testé.

On observe, comme prévu, que la proportion d'exploration diminue à mesure que le nombre de points échantillonnés augmente. A noter qu'il y a encore quelques explorations même après un nombre élevé d'expériences ; en effet, il vaut mieux continuer à explorer pour éviter de tomber sur un maximum local.
Nous avons ensuite testé la sensibilité de notre approche à son initialisation. En effet, le nombre d'expériences M nécessaires pour atteindre 80% de la valeur optimale dépend de la pertinence des points aléatoires initiaux. Nous avons répété la simulation 100 fois où à chaque simulation nous avons 5 points initiaux aléatoires avec $y < 0,1$ (c'est-à-dire que nous avons choisi des points non pertinents). Nous avons évalué que le nombre d'itérations nécessaires pour atteindre 80% de la valeur optimale est $M = 20 \pm 12.5$. Cette variabilité est due au fait que l'algorithme est sensible aux points initiaux, d'autant plus quand le nombre de points initiaux est faible.

Fonction sinus avec $N > 3$ variables d'entrée Pour nous rapprocher d'un problème du monde réel, nous avons testé notre approche avec plus de variables. Pour cela, nous considérons la fonction objectif suivante avec des $N \geq 3$ variables d'entrée :

$$f_N(\boldsymbol{x}) = \left| \prod_{k=1}^{N} \sin(i\pi x_k) \right| \tag{14}$$

avec $\boldsymbol{x} = (x_1, x_2..., x_N)$ variables continues, discrètes ou catégorielles. Les valeurs de y sont comprises entre 0 et 1. Plus précisément, nous étudions le cas de variables discrètes d'entrée $3 \leq N \leq 10$ (avec 5 valeurs différentes chacune). Ceci simule une expérience d'optimisation de composition dont les valeurs sont très contraintes.
Nous utilisons ce dernier jeu de données jouet pour comparer notre approche avec l'optimisation bayésienne (OB). L'OB diffère de notre algorithme flou pour les étapes suivantes [1, 20] :

— Pour l'OB, nous évaluons la fonction de substitution à l'aide du processus gaussien (GP) ou de l'algorithme de Parzen structuré arborescent (TPE), pour modéliser la fonction objectif avec une valeur moyenne m et une dispersion σ. Cet algorithme rend l'OB non déterministe. Pour notre algorithme flou, la fonction surrogate est obtenue à partir de l'algorithme de régression basé sur le clustering flou décrit en partie 3.

— Pour l'OB, le compromis entre exploitation et exploration est modélisé en utilisant la fonction d'acquisition "Expected Improvement" : $EI(\boldsymbol{x}) = (m(\boldsymbol{x}) - f(\boldsymbol{x}^*))\phi\left(\frac{m(\boldsymbol{x})-f(\boldsymbol{x}^*)}{\sigma(\boldsymbol{x})}\right) + \sigma\Phi\left(\frac{m(\boldsymbol{x})-f(\boldsymbol{x}^*)}{\sigma(\boldsymbol{x})}\right)$, avec ϕ/Φ la densité de probabilité/fonction de partition de la distribution normale et $f(\boldsymbol{x}^*)$ la meilleure valeur y obtenue jusqu'à présent. Pour notre algorithme flou, ce compromis exploitation/exploration est modélisé à travers le score d'échantillonnage $S(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) + \lambda\, d(\boldsymbol{x})$.

De manière analogue à notre algorithme, l'OB est répétée sur un certain nombre d'itérations. Chaque boucle fournit des informations supplémentaires jusqu'à atteindre une valeur optimale. L'algorithme TPE est efficace en termes de calculs et bien adapté aux problèmes d'optimisation de grande dimension avec une fonction objectif coûteuse [21]. De plus, il est bien adapté aux problèmes d'optimisation impliquant des variables discrètes et catégorielles, ainsi que des variables continues [22].
Pour les deux algorithmes, nous évaluons le nombre d'itérations M nécessaires pour atteindre 80% de la valeur optimale pour un nombre N donné de variables d'entrée et pour 5 points initiaux donnés (voir Fig. 10). Dans le cas de l'OB, en raison du caractère aléatoire du calcul de la fonction surrogate, nous avons dû répéter la simulation plusieurs fois pour obtenir une valeur moyenne. Nous observons que globalement notre algorithme flou donne un meilleur résultat que l'OB, il converge avec moins d'itérations :

$$\forall N, M_{fuzzy} < M_{BO}. \tag{15}$$

De plus, l'OB est une méthode non déterministe et on observe que la disparité de convergence est assez importante (voir les barres d'écart type élevées en bleu sur la figure).

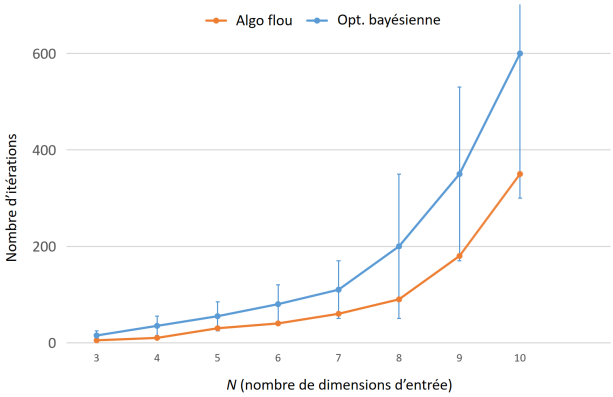

FIGURE 10 – Nombre d'itérations nécessaires pour atteindre 80% de la valeur optimale - Cas avec N variables discrètes.

6.3 Jeu de données réel

Nous avons ensuite testé notre approche sur un ensemble de données réelles du UC Irvine Machine Learning Repository appelé "Concrete Compressive Strength" [1].
Il comprend 8 variables d'entrée et le but est de maximiser la variable de sortie normalisée "Concrete compressive strength (MPa)". Cet ensemble de données comprend 1030 instances. En utilisant un processus de validation croisée (avec 80% de l'ensemble d'entraînement et 20% de l'ensemble de tests), nos simulations sur cet ensemble de données ont montré que la fonction de substitution de notre algorithme flou conduit à une meilleure prédiction de régression que le processus gaussien d'optimisation bayésienne : $RMSE = 8,7 \pm 2,9$ pour notre algorithme flou, et $RMSE = 15,1 \pm 3,6$ pour le processus gaussien. En utilisant notre substitut décrit dans la partie 5, nous obtenons $RMSE = 10.7 \pm 4.9$; ce résultat est moins bon que notre algorithme, soulignant la nécessité d'utiliser les règles N-dimensionnelles décrites dans la partie 3.2.
Pour déterminer le nombre d'expériences pour atteindre un point optimal, nous avons effectué le processus suivant : nous avons choisi 5 points aléatoires parmi les 1030 instances avec une mauvaise valeur de sortie $y < 0,1$; une expérience consiste ici à prendre un point parmi les instances choisies par l'algorithme et on souhaite converger rapidement vers un point optimal. La simulation complète est répétée plusieurs fois pour obtenir une valeur moyenne et un écart type. Le nombre d'expériences nécessaires pour atteindre 80% de la valeur optimale de y est $M = 9 \pm 6,7$. Qualitativement, notre algorithme est particulièrement utile lorsqu'il s'agit de données expérimentales. En effet, l'explicabilité est nécessaire pour comprendre pourquoi un point/une région précis est proposé par l'algorithme. Le dé-

1. https://archive.ics.uci.edu/dataset/165/concrete+compressive+strengt

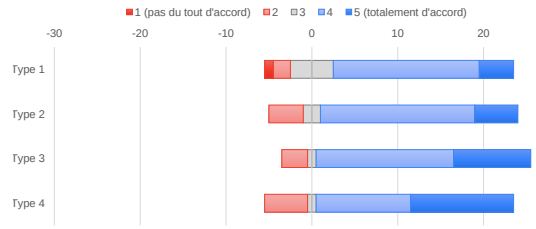

FIGURE 11 – Réponses à la question “Êtes-vous satisfait de la prochaine expérience proposée par l'algorithme ?”

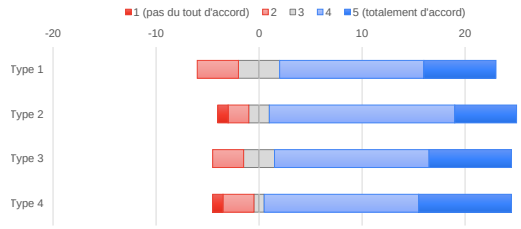

FIGURE 12 – Réponses à la question “Faites-vous confiance au choix de l'algorithme ?”

terminisme est également très important puisque l'expérimentateur ne souhaite pas se voir proposer un point différent à chaque fois qu'il exécute l'algorithme. De plus, des contraintes peuvent être prises en compte sur la base de connaissances expérimentales et théoriques.

7 Considération de l'utilisateur final

Nous avons évalué l'interaction avec les utilisateurs finaux à l'aide d'un questionnaire (évaluation basée sur l'humain). Le panel est constitué de 29 personnes travaillant dans le domaine de la science des matériaux, allant des chercheurs académiques aux chercheurs industriels, âgées de 22 à 62 ans. Pour recueillir leurs avis, nous avons utilisé une échelle de Likert en 5 points, allant de totalement en désaccord à totalement d'accord.

Le questionnaire décrit une situation basée sur un mélange de 3 composés (x_1, x_2, x_3) et une propriété (y) qui va de 0 à 1. Sur un diagramme ternaire, 17 expériences précédentes sont représentées avec les valeurs respectives de la propriété. Nous avons demandé au panel de comparer 4 résultats différents :

- Type 1 : l'algorithme donne exactement les valeurs suivantes pour x_1, x_2, x_3, comme dans l'équation. 9 ;
- Type 2 : idem Type 1 avec une explication (ex : “exploiter une zone déjà explorée et qui a donné de bons résultats”) ;
- Type 3 : l'algorithme donne une région d'intérêt comme dans l'équation 10 ;
- Type 4 : identique au Type 3 et avec la même explication que dans le Type 2.

La figure 11 montre les réponses à la question : “êtes-vous satisfait de la ou des prochaines expériences proposées par l'algorithme ?”. Les résultats sont majoritairement positifs, mais le troisième type de production semble avoir les meilleures notes (c'est-à-dire une région sans explication). De plus, la Figure 12 montre les réponses à la question : “faites-vous confiance au choix de l'algorithme ?”. Les deux types qui n'apportent aucune explication ont des résultats moins positifs. Cependant, les deux types qui fournissent une explication aux utilisateurs ont des résultats plus positifs, mais ils ont également un panéliste qui est totalement en désaccord. Nous avons regardé de plus près ses réponses et elles sont parfois contradictoires : cela peut être considéré comme une valeur aberrante.

Nous avons demandé aux panélistes quelle est leur sortie préférée : 12 d'entre eux répondent au Type 4, 8 répondent au Type 2, 5 pour le type 3 et enfin 4 pour le Type 1. Ainsi, les deux sorties préférées sont accompagnées d'explications.

Il est important de mentionner que nous avons demandé aux panélistes s'ils avaient peur de l'intelligence artificielle dans leur travail : 7 panélistes sont tout à fait d'accord et 10 d'entre eux sont d'accord. 9 d'entre eux sont restés neutres, ce qui signifie que seuls 3 d'entre eux n'ont pas peur.

Les résultats confirment ce que nous attendions : les utilisateurs finaux préfèrent avoir un choix et une explication, ce qui signifie que la méthode finale doit fournir la région de la prochaine expérience et une explication de la recommandation. Dans un commentaire, un panéliste a écrit qu'il préférait avoir plusieurs expériences possibles plutôt qu'une seule, mais puisque l'algorithme a choisi une expérience centrale, son choix serait le même. Il souligne l'importance de prendre en compte les préférences humaines dans de tels outils pour accroître leur acceptabilité.

8 Conclusion et perspectives

En conclusion, nous résumons les avantages de notre approche en fonction des résultats obtenus. Notre algorithme est transparent en conséquence directe du caractère interprétable de ses paramètres, de la prédominance d'un cluster dans chaque région de l'espace d'entrée-sortie, de la simplicité et de la nature linguistique de ses règles floues ; il est déterministe, c'est-à-dire qu'il converge vers les mêmes valeurs à chaque exécution ; il est nettement plus rapide que les approches méta-heuristiques telles que les algorithmes évolutionnaires ; il est robuste au surentraînement et résilient au bruit grâce à la contribution fusionnée des clusters ; enfin, des contraintes issues de la littérature peuvent être appliquées pour réduire l'espace de recherche d'entrée. Cette approche a le mérite d'être interprétable, intuitive, et peut être d'une réelle aide aux expérimentateurs.

L'originalité de notre algorithme inclut la transformation de clusters flous multidimensionnels en ensembles flous 1D interprétables, et la définition d'une fonction de score d'acquisition/échantillonnage à partir des variables $\hat{y}$ (valeur de sortie prédite) et d (distance à le point échantillonné le plus proche).

Les possibilités d'amélioration incluent la vitesse de calcul dans le cas de grande dimension, une meilleure détection

des extrema locaux et une optimisation multi-objectifs.

Références

[1] S. GREENHILL et al. “Bayesian optimization for adaptive experimental design : A review”. In : *IEEE access* 8 (2020), p. 13937-13948.

[2] D.S. BEM et al. “Combinatorial experimental design using the optimal-coverage algorithm”. In : *Experimental Design for Combinatorial and High Throughput Materials Development* (2003).

[3] S. TAKAMOTO et al. “Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements”. In : *Nature Communications* 13.2991 (2022).

[4] A. PAGLIARO et P. SANGIORGI. “AI in Experiments : Present Status and Future Prospects”. In : *Applied Sciences* 13.10415 (2023).

[5] B. CAO et al. “How To Optimize Materials and Devices via Design of Experiments and Machine Learning : Demonstration Using Organic Photovoltaics”. In : *ACS Nano* 12.8 (2018), p. 7434-7444.

[6] F. REN et al. “Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments”. In : *Science Advances* 4.4 (2018), eaaq1566.

[7] D. XUE et al. “Accelerated search for materials with targeted properties by adaptive design”. In : *Nature communications* 7.1 (2016), p. 1-9.

[8] S.L. CHIU. “Fuzzy model identification based on cluster estimation”. In : *Journal of Intelligent & fuzzy systems* 2.3 (1994), p. 267-278.

[9] K. WIKTOROWICZ. “RFIS : regression-based fuzzy inference system”. In : *Neural Computing and Applications* 34 (juill. 2022).

[10] J. VIAÑA et al. “Explainable fuzzy cluster-based regression algorithm with gradient descent learning”. In : *Complex Engineering Systems* (2022).

[11] J.C. BEZDEK, R. EHRLICH et W. FULL. “FCM : The fuzzy c-means clustering algorithm”. In : *Computers & Geosciences* 10.2 (1984), p. 191-203. ISSN : 0098-3004.

[12] I.B. TÜRKŞEN. “A review of developments from fuzzy rule bases to fuzzy functions”. In : *2012 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)*. 2012, p. 1-5.

[13] M.-J.e LESOT, M. RIFQI et B. BOUCHON-MEUNIER. “Fuzzy Prototypes : From a Cognitive View to a Machine Learning Principle”. In : *Fuzzy Sets and Their Extensions : Representation, Aggregation and Models*. Sous la dir. d’Humberto BUSTINCE, Francisco HERRERA et Javier MONTERO. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 431-452. ISBN : 978-3-540-73723-0.

[14] G. TSEKOURAS et al. “A hierarchical fuzzy-clustering approach to fuzzy modeling”. In : *Fuzzy sets and systems* 150.2 (2005), p. 245-266.

[15] S. GUILLAUME et B. CHARNOMORDIC. “Generating an interpretable family of fuzzy partitions from data”. In : *IEEE Transactions on Fuzzy Systems* 12.3 (2004), p. 324-335.

[16] S. MANEEWONGVATANA et D.M. MOUNT. “Data Structures, Near Neighbor Searches, and Methodology”. In : American Mathematical Society, 2002. Chap. Analysis of approximate nearest neighbor searching with clustered point sets.

[17] L.-J. LI et Y.-L. LIANG. “A Hierarchical Fuzzy Clustering Algorithm”. In : *International Conference on Computer Application and System Modeling* (2010).

[18] A.R. CONN, N.I.M. GOULD et P.L. TOINT. “Trust-Region Methods”. In : *MPS-SIAM Series on Optimization 1. SIAM and MPS, Philadelphia* (2000).

[19] G.E. TSEKOURAS et al. “A fuzzy clustering-based algorithm for fuzzy modeling.” In : *WSEAS Transactions on Systems* 3.5 (2004), p. 1958-1963.

[20] S. WATANABE. “Tree-structured Parzen estimator : Understanding its algorithm components and their roles for better empirical performance”. In : *arXiv preprint arXiv :2304.11127* (2023).

[21] J. BERGSTRA, D. YAMINS et D.D. COX. “Making a Science of Model Search : Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures”. In : *Proc. of the 30th International Conference on Machine Learning* (2013).

[22] T. AKIBA et al. “Optuna : A next-generation hyperparameter optimization framework”. In : *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, p. 2623-2631.

Index des notations

- N nombre de dimensions d’entrée
- $\boldsymbol{x}_s$ points déjà échantillonnés
- $\boldsymbol{x}$ point de l’espace d’entrée
- $\boldsymbol{x}_{next}$ prochain point testé
- f ensemble flou
- c cluster
- C ensemble des clusters
- $\boldsymbol{m}_c$ centre du cluster c
- n_c nombre de clusters
- μ_c degré d’appartenance au cluster c
- R règle floue
- $\boldsymbol{\alpha}_c, \beta_c$ coefficients de régression relatifs au cluster c
- $\hat{y}_c$ sortie prévue pour une entrée $\boldsymbol{x}$, par rapport au cluster c
- $\hat{y}$ sortie globale prédite pour une entrée $\boldsymbol{x}$
- M nombre d’itérations nécessaires pour atteindre un point optimal

Session 3 : Apprentissage Automatique

Adaptation de Yolov8 pour la détection d'objets avec peu d'exemples

G. Fourret[1,3], C. Fiorio[1], G. Subsol[1], M. Chaumont[1,2]

[1] Équipe ICAR, LIRMM, Univ. Montpellier, CNRS, France
[2] Univ. Nîmes, France
[3] Drone Geofencing, Nîmes, France

{guillaume.fourret, fiorio, gerard.subsol, marc.chaumont}@lirmm.fr

Résumé

Les réseaux récents pour la détection d'objets obtiennent d'excellentes performances quand ils sont entraînés sur de grandes bases de données, mais ont toujours des difficultés pour apprendre un nouvel objet avec peu d'exemples. Les méthodes de ce domaine utilisant plutôt des architectures lourdes, nous avons décidé d'adapter les modules présentés dans DeFRCN dans l'architecture plus récente et rapide de Yolov8. Nous montrons ici l'impact de cette modification sur le benchmark MSCOCO, et finalement parlons des biais de cette méthode.

Mots-clés

Yolov8, Détection d'objets en peu d'exemples, Apprentissage par transfert

Abstract

Recent networks for object detection obtain excellent performance when trained on large databases but still have difficulties learning a new object with few examples. The methods of this domain using rather heavy architectures, we have decided to adapt the modules presented in DeFRCN into the newer and faster architecture of Yolov8. We show here the impact of this modification on the MSCOCO bechmark, and finally discuss about the biases existing in this method.

Keywords

Yolov8, Few-Shot Object-Detection, Transfer Learning

1 Introduction

Les progrès récents en architectures et méthodes d'entraînement des réseaux neuronaux ont grandement amélioré les performances en vision par ordinateur, notamment en classification, segmentation et détection d'objets (définie par la localisation et classification). La plupart des modèles sont entraînés de manière supervisée sur de vastes ensembles de données annotées, souvent difficiles et coûteux à obtenir.
Les méthodes d'apprentissage avec peu d'exemples [5] ("Few-Shot Object Detection", FSOD) ont émergé pour répondre à ce problème. Deux paradigmes de ce domaine sont le méta-apprentissage [11], qui simule des entraînements avec peu d'exemples et l'apprentissage par transfert [10], qui vise à acquérir de bonnes capacités de représentation sur des ensembles de données volumineux pour les transférer à de nouveaux objets avec peu d'exemples.
Un défi notable dans ce dernier est la capacité d'un détecteur à séparer la localisation et la classification, comme présenté dans DeFRCN [3] avec leur couche de découplage de gradient (GDL). Pour améliorer la classification, un second module, le PCB, utilise des vecteurs supports (prototypes) pré-calculés. Cependant, cette méthode est spécifique à l'architecture en deux étapes du Faster R-CNN [7]. Des détecteurs plus récents en une seule étape, tels que ceux de la famille YOLO, sont apparus offrant des performances et une vitesse d'inférence supérieure. Dans cet article, nous montrons comment nous avons intégré ces deux modules dans l'architecture de Yolov8 [2] pour améliorer ses performances avec peu d'exemples d'entraînement.

2 État de l'art

2.1 Méthodes et paradigmes du FSOD

Toutes les approches suivent un schéma d'entraînement en 2 étapes. Une première base de données est utilisée pour le pré-entraînement sur les classes de base. Une deuxième base de données sert ensuite pour l'apprentissage des nouvelles classes et est constituée de K exemples ("K-shots") de *toutes* les classes.
Le méta-apprentissage [11] simule des tâches FSOD en sous-échantillonnant K-shots des classes de base lors du pré-entraînement. La plupart des méthodes utilisent une branche en parallèle d'un Faster R-CNN pour créer des vecteurs supports/prototypes représentant chaque classe. Ces techniques diffèrent principalement par l'endroit où ces vecteurs sont agrégés dans le Faster R-CNN (avant [11] ou [6] après le "Region Proposal Network" RPN) et par l'opération d'agrégation utilisée (produit simple, attention...).
L'apprentissage par transfert a émergé récemment comme un paradigme plus performant. Dans [10], un simple apprentissage (en gelant l'encodeur) d'un Faster R-CNN pré-entraîné a surpassé les méthodes de méta-apprentissage. De plus, ces auteurs ont établi les méthodes actuelles d'évaluation pour les méthodes FSOD, en particulier sur MSCOCO [4]. Le Faster R-CNN a été réutilisé dans DeFRCN [3] et

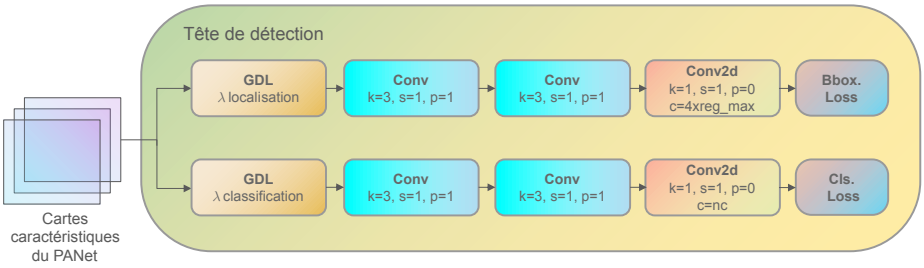

FIGURE 1 – Schéma d'architecture d'une tête de détection de Yolov8 avec les couches GDL.

dans DCFS [1] pour sa capacité à séparer la classification et la localisation. L'apprentissage de ces deux tâches exige que le réseau apprenne des caractéristiques spécifiques aux classes pour la classification, ainsi que d'autres invariantes pour la localisation. En FSOD, ce problème est amplifié en l'absence d'un nombre suffisant d'exemples pour trouver un compromis. Le module GDL de DeFRCN résout ce problème en ajustant le gradient provenant du RPN et de la tête de classification R-CNN. De plus, leur module PCB compare des prototypes pré-calculés aux prédictions du réseau pendant l'inférence pour ajuster les scores de classification.

2.2 Yolov8

Yolov8 est la dernière version des algorithmes YOLO et introduit un changement de paradigme. En effet, Yolov8 est un modèle n'utilisant pas d'ancres pré-calculées pour ses détections, ce qui le rend facilement adaptable à d'autres tâches. Le but de ces modèles sans ancres est d'effectuer des prédictions denses (i.e : pour chaque pixel de leurs espaces latents). D'une manière générale, Yolov8 est inspiré de l'architecture présentée dans [12]. Pour obtenir des représentations à plusieurs niveaux, les auteurs utilisent un encodeur connecté à un "Feature Pyramid Network" [9] et, pour Yolov8, un "Path Aggregation Network" [8]. Puis, Yolov8 utilise pour chacun de ces 3 niveaux de résolution une tête de détection composée de deux branches pour la localisation et la classification. Pour les pertes, la localisation utilise une variante de la Focal Loss et de l'IoU, tandis que la classification utilise une entropie croisée.
Yolov8 donne certes de très bons résultats dans un cadre classique, mais n'est pas adapté au problème du FSOD.

3 Intégration du GDL et PCB dans Yolov8

3.1 Gradient Decoupling Layer

Depuis Yolov1, la prédiction de la classification et de la localisation se fait à partir des mêmes poids, ce qui rend un découplage impossible. Avec son changement d'architecture, Yolov8 utilise des branches distinctes en parallèle. La branche de localisation agit alors comme un détecteur d'objets "générique" qui propose des boîtes englobantes, tandis que la branche de classification sort des "heatmaps" pour chacune des classes. Cette nouvelle architecture rend possible l'intégration du module GDL en ajoutant ces couches de découplage au début de chacune des branches, comme illustré dans la figure 1, la rétropropagation dans Yolov8 devenant alors :

$$\theta_{PANet} \leftarrow \theta_{PANet} - \gamma(\lambda_{loc}\frac{\partial \mathcal{L}_{loc}}{\partial \theta_{PANet}} + \lambda_{cls}\frac{\partial \mathcal{L}_{cls}}{\partial \theta_{PANet}}) \tag{1}$$

Où θ_{PANet} représente les paramètres du backbone et du FPN+PANet de Yolov8, $\mathcal{L}_{loc}$, $\mathcal{L}_{cls}$ et λ_{loc}, λ_{cls} la perte et les valeurs de découplage de gradients sur les branches de localisation et de classification, et γ le learning rate.

3.2 Prototypical Calibration Block

Nous calculons d'abord les prototypes en donnant les K-shots de chaque classe à un extracteur de caractéristiques pré-entraîné afin d'obtenir des vecteurs. Ensuite, le PCB calcule le prototype p_c pour chaque classe en moyennant ces vecteurs, formant la banque de prototypes $P = \{p_c\}$.
Puis dans Yolov8, comme illustré dans la figure 2, nous prenons chaque boîte englobante proposée par la branche de localisation en entrée d'un "RoI Align" afin d'extraire de la feature map de l'image en inférence des représentations de taille fixe f_{bbox}. Enfin, pour chaque boîte englobante, on calcule la similarité cosinus entre sa représentation et les prototypes de chaque classe pour obtenir son score de classification d'après le module PCB :

$$PCB_{bbox} = \frac{f_{bbox} \cdot p_c}{\|f_{bbox}\|\|p_c\|}, p_c \in P \tag{2}$$

Pour finir, nous fusionnons les scores de classification de toutes les boîtes englobantes du PCB et de Yolov8 C_{yolo} avec une somme pondérée par α en profondeur sur les heatmaps pour obtenir le score de classification finale C_{final} :

$$C_{final} = \alpha \cdot C_{yolo} + (1 - \alpha) \cdot PCB \tag{3}$$

4 Expérimentations

Nous avons évalué l'impact de ces modules dans Yolov8 sur MSCOCO adapté au FSOD, comme décrit dans [10]. Sur les 80 classes, 60 sont les classes de base, tandis que les 20 dernières sont les nouvelles. Pour l'entraînement, toutes les images de MSCOCO ont été utilisées, à l'exception de 5000 réservées pour la validation. Les métriques utilisées sont bAP50,bAP,nAP50,nAP qui correspondent au mAP50 et mAP pour les classes de base et les nouvelles classes.
Module GDL. Nous avons tout d'abord effectué plusieurs pré-entraînements avec le modèle Yolov8m en modifiant les valeurs de découplage. Cependant, dans ce cas classique avec beaucoup de données, nous n'avons pas observé d'amélioration comme dans DeFRCN pour le Faster

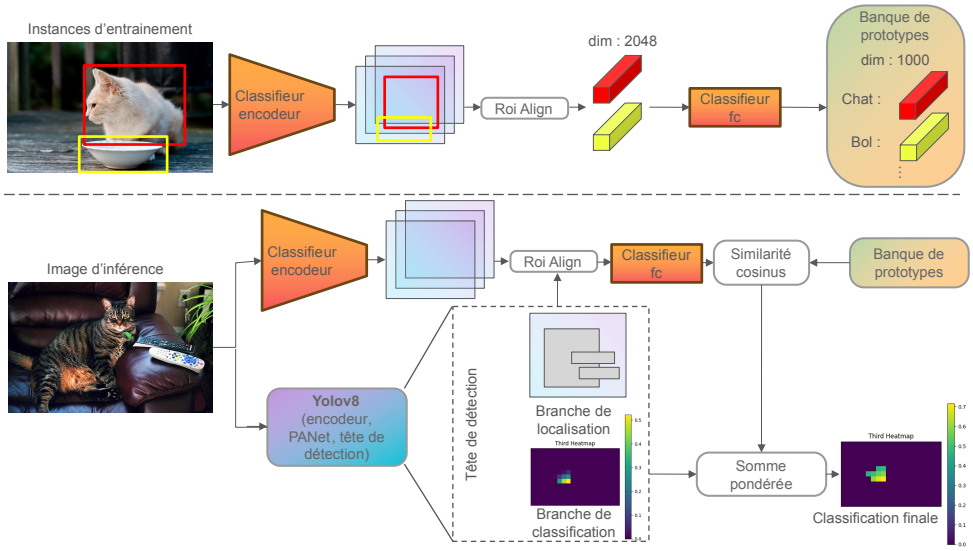

FIGURE 2 – Illustration du module PCB pour une des trois têtes de détection et seulement la heatmap pour la classe *Chat*. "*Classifieur fc*" dénote la dernière couche complètement connectée du classifieur pré-entraîné sur ImageNet.

	1-shot				*2-shot*			
	bAP50	bAP	nAP50	nAP	bAP50	bAP	nAP50	nAP
Yolov8m_vanilla	13.9	9.1	3.9	2.3	12.9	8.9	6.3	3.7
Yolov8m_GDL	54.7	39.0	10.4	6.9	49.0	34.7	12.0	7.4
DeFRCN	48.9	32.0	8.8	5.1	50.4	32.9	16.8	9.6
	3-shot				*5-shot*			
	bAP50	bAP	nAP50	nAP	bAP50	bAP	nAP50	nAP
Yolov8m_vanilla	19.1	13.4	9.3	5.6	19.3	13.6	11.7	7.4
Yolov8m_GDL	56.0	40.7	15.1	10.2	53.2	38.2	21.3	14.0
DeFRCN	50.6	33.1	21.9	12.3	51.5	33.6	26.1	14.2
	10-shot				*30-shot*			
	bAP50	bAP	nAP50	nAP	bAP50	bAP	nAP50	nAP
Yolov8m_vanilla	19.3	13.6	15.5	10.0	23.6	16.9	23.2	15.7
Yolov8m_GDL	50.0	35.1	26.2	16.9	52.1	36.5	31.8	21.1
DeFRCN	53.3	34.6	32.2	17.3	53.6	34.9	38.3	21.4

TABLE 1 – Comparaison entre Yolov8 vanilla, Yolov8 avec module GDL, et DeFRCN avec GDL uniquement.

R-CNN, les meilleurs modèles étant ceux rétro-propageant également la localisation. Tous ces pré-entraînements obtiennent cependant de meilleurs scores que DeFRCN étant donné l'architecture plus récente de Yolov8 (+2.1 bAP50 sans localisation, +4.3 bAP50 avec).

Nous avons effectué ensuite la phase d'apprentissage K-shots en testant pour chacun des pré-entraînements de nouvelles valeurs de découplage. Le meilleur résultat que nous ayons obtenu fut en utilisant pour le pré-entraînement $\lambda_{loc} = 0.25$ et $\lambda_{cls} = 0.75$, et pour le finetuning $\lambda_{loc} = 0.1$ et $\lambda_{cls} = 0.1$. Les résultats du Yolov8 résultant, présentés dans le tableau 1, montrent une amélioration par rapport à Yolov8 classique et des scores presque équivalent à DeFRCN en nAP pour le 5,10, et 30-shots, tout en ayant 2 fois moins de paramètres (25M vs 51M).

Nous avons également testé d'autres tailles de Yolov8. Le modèle s donne des performances inférieures (-3.4 nAP50), et le modèle x n'apporte pas d'améliorations notables.

Module PCB. Comme dans DeFRCN, nous avons utilisé le modèle Resnet101 pré-entrainé sur ImageNet comme extracteur de caractéristiques. Nous avons effectué des tests avec plusieurs valeurs de α. De plus, nous appliquons ce module uniquement lorsque Yolov8 prédit une nouvelle classe afin de ne pas potentiellement détériorer les classes de base qui sont déjà bien apprises par le réseau. Enfin, nous appliquons le module uniquement lorsque la confiance de Yolov8 se situe entre une borne inférieure c_{min} et une borne supérieure c_{max}. Le module PCB intégré à Yolov8 nous apporte un léger gain, détaillé dans le tableau 2, en utilisant les paramètres $\alpha = 0.5$, $c_{min} = 0.05$ et $c_{max} = 1$.

5 Discussions

L'utilisation des modules GDL et PCB peut donc apporter un gain pour Yolov8 dans le contexte du FSOD. Cependant, nous voulons remettre en perspective nos résultats en analysant certains biais que nous pensons qu'il existe avec cette méthode (et d'autres dans la littérature).

1-shot		*2-shot*		*3-shot*	
nAP50	nAP	nAP50	nAP	nAP50	nAP
+0.1	+0.1	+0.7	+0.4	+0.8	+0.6
5-shot		*10-shot*		*30-shot*	
nAP50	nAP	nAP50	nAP	nAP50	nAP
+1.7	+1.0	+1.7	+1.0	+1.6	+1.0

TABLE 2 – Gain du module PCB sur les nouvelles classes.

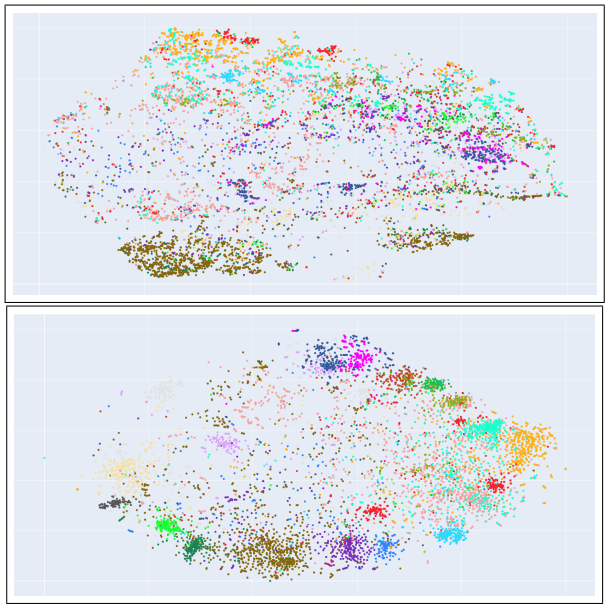

FIGURE 3 – Visualisation t-SNE des représentations des nouvelles classes par *a*) Yolov8 pré-entraîné sur les classes de base *b*) Resnet101 pré-entraîné sur ImageNet.

5.1 Utilisation d'un extracteur de caractéristiques pré-entrainé

Le module PCB, comme beaucoup de méthodes utilisant des prototypes pré-calculés, utilise un extracteur de caractéristiques pré-entraîné sur ImageNet. Nous aimerions souligner ici que l'utilisation de ces méthodes semble biaisée dans le cas de l'apprentissage few-shot. Bien qu'aucune des classes de ImageNet n'ait exactement le même nom que celles des nouvelles classes de MSCOCO, on peut trouver des objets très similaires sémantiquement. Par exemple pour les animaux, les classes *Cat*, *Dog*, sont de nouvelles classes à apprendre, or certaines classes présentent dans ImageNet partagent beaucoup de traits avec celles-ci (*Siamese Cat*, *Persian Cat*, *Shetland Sheepdog* ...). On peut d'ailleurs observer l'impact de ce pré-entraînement en comparant les représentations de ces nouvelles classes figure 3.

5.2 Suivi des performances sur la validation

Pendant l'apprentissage des nouvelles classes sur MSCOCO, un suivi des performances du détecteur est calculé sur l'ensemble de validation à chaque époque afin de pouvoir obtenir les meilleurs poids. Cela nécessite donc beaucoup d'exemples de ces nouvelles classes dans l'ensemble de validation ce qui n'est pas vraiment crédible dans un cas d'application réel du FSOD où les seuls exemples sont ceux de l'entraînement. Pour essayer de quantifier cet impact, nous avons comparé pour un même Yolov8 les différences avec et sans suivi (l'entraînement s'arrêtant au bout de 180 époques). Sans ce suivi peu réaliste, nous obtenons une différence notable de -9.7 bAP50 et -2.4 nAP50.

Remerciements

Nous remercions l'Association Nationale de la Recherche et de la Technologie ainsi que Drone Geofencing pour le financement de la thèse CIFRE.

Références

[1] Bin-Bin Gao et al. Decoupling classifier for boosting few-shot object detection and instance segmentation. *Advances in Neural Information Processing Systems*, 35 :18640–18652, 2022.

[2] Glenn Jocher et al. YOLO by Ultralytics, January 2023.

[3] Limeng Qiao et al. Defrcn : Decoupled faster r-cnn for few-shot object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8681–8690, 2021.

[4] Michael Maire et al. Microsoft coco : Common objects in context. pages 740–755. Springer International Publishing, 2014.

[5] Mona Köhler et al. Few-shot object detection : a comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[6] Qi Fan et al. Few-shot object detection with attention-rpn and multi-relation detector. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4013–4022, 2020.

[7] Shaoqing Ren et al. Faster r-cnn : Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[8] Shu Liu et al. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Tsung-Yi Lin et al. Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[10] Xin Wang et al. Frustratingly simple few-shot object detection. volume 119, pages 9919–9928. PMLR, 4 2020.

[11] Xiongwei Wu et al. Meta-rcnn : Meta learning for few-shot object detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1679–1687, 2020.

[12] Zhi Tian et al. Fcos : Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

Plateformisation de l'apprentissage machine en épidémiologie : enjeux philosophiques

É. Pardoux[1], T. Guyet[2]

[1] CNRS & ENS Lyon, IHRIM & MFO
[2] INRIA Lyon

eric.pardoux@ens-lyon.fr ; thomas.guyet@inria.fr

Résumé

L'épidémiologie se numérise massivement grâce la disponibilité croissante des données de santé. Elle tend à utiliser de plus en plus d'outils informatiques pour automatiser la réalisation d'études. Les enjeux éthiques de l'automatisation en santé sont bien connus, néanmoins leur résolution au sein du système même est complexe. À défaut de rendre un outil éthique, on s'interroge sur la possibilité et les moyens de systématiser une réflexion éthique dans la conception d'études épidémiologiques utilisant des données de santé. L'ambition de ce projet est de fournir une plateforme qui facilite la reproduction des études épidémiologiques tout en soutenant la réflexion éthique. Au travers de cet article prospectif, nous mêlons des réflexions techniques et philosophiques sur une telle plateforme.

Mots-clés

Données de santé, Épidémiologie, Philosophie des techniques, FAIRification, Éthique de l'IA

Abstract

Epidemiology is becoming massively digitized, thanks to the growing availability of health data. It tends to use more and more computerised tools to automate the conduct of studies. The ethical challenges of automation in healthcare are well known, but resolving them within the system itself is complex. Making the system itself ethical appears to be difficult, albeit impossible : the question is therefore whether and how it can instead foster ethical reflection in the design of epidemiological studies using health data. The ambition of this project is to provide a platform that facilitates study replication and supports ethical reflection. In this prospective article, we combine technical and philosophical reflections on such a platform.

Keywords

Health data, Epidemiology, Philosophy of technology, FAIRification, AI ethics

1 Introduction

Les études épidémiologiques sont fondamentales pour la médecine fondée sur les preuves. Elles permettent de mettre en évidence des liens entre des causes probables et des effets observés au cours de maladies et de pathologies. L'épidémiologie se numérise massivement ces dernières années grâce la disponibilité croissante des données de santé, ainsi que des progrès dans les moyens de calcul – au premier rang desquels l'intelligence artificielle (IA) et plus particulièrement le *machine learning* (ML).

Cette introduction de nouvelles techniques de traitement de données est annoncée comme révolutionnaire (*1*), que ce soit au niveau de l'identification des risques pour les personnes, de l'amélioration d'interventions ciblées ou encore de l'identification de nouveaux motifs précurseurs de pathologies dans les parcours de soin. Néanmoins, cette numérisation croissante n'efface pas les problèmes éthiques et scientifiques de l'épidémiologie, voire les augmente (*2*). Idéalement toute étude épidémiologique se doit d'être transparente – notamment quant aux données employées et aux traitements menés – et reproductible. Ceci tout en visant à maximiser les bénéfices pour la santé publique et à minimiser les éventuels nuisances, notamment envers des groupes particuliers au sein des populations, dont il s'agit à la fois de s'assurer la représentativité en fonction des problèmes adressés et leur consentement au réemploi de leurs données de santé. Concevoir une étude épidémiologique représente donc une réflexion dynamique plus large qu'un simple raisonnement statique sur des données populationnelles de santé [1]. La reproduction de résultats d'études épidémiologiques est rendue à l'heure actuelle difficile du fait de l'accès non garanti aux données ou encore de l'absence récurrente du contexte et des méta-données employées. Les pipelines de traitement de données peuvent ainsi s'avérer complexes à reconstruire indépendamment. Développer une plateforme centralisant ces éléments faciliterait la reproduction des études épidémiologiques.

L'ambition de notre projet est de fournir une plateforme qui permette de soutenir la réflexion épidémiologique autour des enjeux éthiques et scientifiques sans pour autant s'y substituer. Pour ce faire, le raisonnement épidémiolo-

1. Nous effectuons ici une distinction entre la réflexion, comprise comme l'activité générique cognitive de pensée à propos d'un sujet et de son contexte, et le raisonnement, compris comme suite d'arguments organisés suivant des principes et règles logiques – soit le sens classique du terme en IA : le contexte se limite aux (méta)données fournies au système.

gique moderne, déjà fortement codifié dans le domaine de la santé publique (*3*), est vu comme un *workflow* scientifique (*4*), *i.e.* une suite d'opération de collecte de données et d'analyses statistiques conduisant à la réponse à une question précise. Il s'agit de traiter ces *workflows* suivant les principes FAIR [2] (*5*), tout en intégrant des outils facilitant une conception éthique des études épidémiologiques à partir des données de santé, au sein d'une seule plateforme.
Ces considérations ouvrent au plan de l'éthique, où elles retrouvent les questions fondamentales de représentativité et d'équité dans le traitement des données. De quelle façon pourraient-elles être incluses dans la plateforme ou au contraire à quel point la responsabilité reste fonction du cadre d'usage ? La FAIRification suffit-elle à répondre à ces enjeux éthiques ? Ou bien le caractère éthique du système algorithmique reste-t-il toujours à construire, avec l'aide de ce dernier et contre ses dérives ?
Face aux enjeux visant à faire se rencontrer des parties prenantes aux connaissances et intérêts variés [3], la question est celle de savoir comment faciliter une réflexion éthique systématique à défaut de rendre le système d'aide à la décision éthique. En d'autres termes, la plateforme doit catalyser et faciliter la réflexion éthique chez l'épidémiologiste, plutôt que d'être éthique en elle-même.

2 Plateforme pour le raisonnement épidémiologique

L'opportunité d'exploiter les données médico-administratives pour mener certaines recherches épidémiologiques (*6*) a conduit à développer l'utilisation des méthodologies de l'analyse de données au sein de ce domaine de recherche.

Exemple 1 *Prenons l'exemple d'une étude sur l'impact sur la survie en fonction du délai de prise en charge des patients pour exérèse pulmonaire suite au diagnostic d'un cancer du poumon. L'approche épidémiologique classique conduirait à collecter pendant plusieurs années des informations sur des patientes et patients qui seront opérées, puis à mener une étude statistique. Dans une approche utilisant des données de santé, nous utilisons des données historisées dans les entrepôts pour identifier des patients d'intérêts et mener l'étude statistique.*
Dans les deux cas, il faut procéder en identifiant des patientes et patients d'intérêt, en collectant des informations sur les délais entre le diagnostic et la chirurgie, et sur le décès éventuel. Enfin, il faut mener une analyse de survie. Dans le cas de l'utilisation des données médico-administratives, toutes ces étapes peuvent être réalisées par des outils informatiques : requêtes SQL et méthodes statistiques.

2. Ces principes visent à rendre les données Faciles à (re)trouver, Accessibles, Interopérables et Réutilisables, pour faciliter leur exploitation autant par les machines que par les individus.

3. On pourra distinguer des nuances notamment entre chercheurs et chercheuses, *data scientists*, épidémiologistes, médecins de santé publique, décideurs et décideuses, la population ou encore les associations de patients et patientes. Ces entités ont toutes un intérêt dans les décisions de santé publique éclairées par l'épidémiologie.

Parfois l'identification de certains évènements d'intérêt nécessite des analyses plus poussées. En particulier, pour récupérer les informations de dates des chirurgies des patients et patientes, il peut être nécessaire d'utiliser des outils d'IA de traitement automatique du langage (TAL) pour analyser les rapports médicaux (7).

Les spécificités de l'approche sur données médico-administratives sont 1) d'exploiter des sources de données massives pour répondre à une multitude de questions de santé publique. C'est le cas en particulier du Système National de Données de Santé (SNDS) ou des données hospitalières qui sont aujourd'hui utilisées dans de nombreuses études ; 2) d'utiliser une collection d'outils de traitement génériques qui sont mobilisés pour pré-traiter, analyser et présenter des données (requêtes à des bases de données, traitements algorithmiques et statistiques, visualisations).
Ces deux caractéristiques rendent possibles une plateformisation concrète de l'épidémiologie, par la réalisation d'une plateforme logicielle au travers de laquelle les épidémiologistes pourraient concevoir et réaliser leurs études.

Exemple 2 *(Ex. 1 cont.) Aujourd'hui, pour conduire l'étude de notre exemple, c'est un ensemble d'outils et de code (e.g., R ou Python) qui sont utilisés de manière ad hoc et sans soucis de cohérence. Le ou la* data scientist *enchaîne manuellement les étapes de son projet (souvent à l'aide de fichiers intermédiaires). Une plateforme vise à structurer et normaliser ces différents outils et processus.*

Les enjeux de la conception d'une telle plateforme sont :

- de permettre aux épidémiologistes, notamment par l'usage d'outils d'analyse avancés d'IA, de tirer au mieux partie des données existantes. Il s'agit de conserver une latitude de programmation et de paramétrisation élevée lors des usages, laissant ainsi le champ des possibles le plus ouvert possible pour l'épidémiologiste. C'est un enjeu d'encapacitation.
- de faciliter la possibilité de partage entre les différentes études épidémiologiques qui seraient entreprises. On peut alors penser en particulier à des enjeux de reproductibilité, ou à la conception de méta-analyses, classiquement menées en épidémiologie pour consolider les résultats des études.
- de favoriser une conception éthique des études épidémiologiques. En effet, on ne peut pas garantir le caractère éthique d'un outil, néanmoins *via* l'intégration native de méthodes conformes aux approches FAIR (*8*) d'une part et l'implémentation d'outils de mesure de l'équité (*9*) ou de respect de la vie privée (*10*) par exemple, nous rendons la plateforme plus apte à soutenir la réflexion éthique de l'épidémiologiste.

Pour mener des réflexions sur la conception d'une telle plateforme, nous nous appuyons sur la philosophie des techniques. En plus des approches FAIR, la démarche poursuivie au sein de ce projet se rapproche du *metadesign* (*11*, *12*), qui consiste pour les concepteurs que nous sommes à

redonner la main sur le produit final du système (les *workflows*) aux posseusses et possesseurs du problème (les épidémiologistes) par le biais d'un outil ouvert et largement configurable.

3 Des études comme des workflows

On peut faire un parallèle entre une étude épidémiologique et un *workflow* computationnel scientifique. De manière générale, un *workflow* scientifique (*5*) est une description formelle de la réalisation d'un objectif scientifique comme un ensemble de tâches et leurs dépendances. De son côté, le *workflow* computationnel (*13*) permet la réalisation (automatique) des étapes d'un *workflow*, chacune de ces étapes étant un traitement computationnel. Dans la plupart des études épidémiologiques menées sur des données médico-administratives, on retrouve systématiquement les mêmes jalons : sélection de patients, transformation des données (*feature engineering*) et analyse des données. Ces étapes correspondent à celles classiques des études épidémiologiques. Elles se distinguent néanmoins par le fait qu'elles sont concrètement des traitements computationnels appliqués sur des données numériques (requêtes SQL, traitements algorithmiques). L'étude épidémiologique complète peut alors être vue comme un *workflow* computationnel.

Le travail de l'épidémiologiste revient alors à concevoir un *workflow* qui répond à la question scientifique d'intérêt. Un premier objectif est de le décharger des contraintes techniques de la manipulation des données pour qu'il se focalise sur la réflexion épidémiologique.

Concrètement, la plateforme proposée se traduit par :

1) la définition d'un **formalisme générique pour la représentation de données**. Nous avons opté pour une représentation des données sous la forme d'une cohorte longitudinale. Un individu du jeu de données est un patient ou une patiente décrite par un ensemble d'attributs statiques et un parcours de soins (ensemble d'évènements datés). La cohorte est ainsi associée à un schéma de description des données (attributs statiques et dynamiques). Le schéma et les données pouvant être annotés par des méta-données, notamment leur provenance. L'utilisation de données non-structurées (documents textuels, images, etc.) sera intégrée comme attribut.
2) la spécification de la notion d'**étape d'un *workflow***, *i.e.* un traitement "élémentaire" d'une cohorte. Dans notre spécification, un traitement transforme une cohorte (spécifiée par schéma $\mathcal{C}$) dans une nouvelle cohorte (spécifiée par nouveau schéma $\mathcal{C}'$) et réalisant également des modifications des données (et des méta-données). [4] Le traitement lui-même est implémenté sous la forme d'un script (*e.g.* Python).
3) un **éditeur de *workflows*** qui permet à l'épidémiologiste lors de l'utilisation de construire de manière interactive le *workflow* de son étude, de l'exécuter et de visualiser les résultats.

4. On met de côté des traitements spécifiques liés à l'import et à l'export des données du *workflow*.

En complément, un ensemble de traitements de base sera fourni pour mettre en place des *workflows* usuels en épidémiologie. Les briques de bases des *workflows* à partir de parcours de soin comprennent (*14*) : le chargement de données, la sélection de patients, la sélection et transformation d'évènements, la détection d'évènements de santé.

Le déploiement d'outils à base de modèle d'IA [5] comme étapes d'un *workflow* fait partie des évolutions attendues de notre plateforme. En particulier, leur utilisation doit permettre d'exploiter des contenus textuels ou d'images. Ils peuvent également servir à identifier des évènements de santé complexes à partir du parcours de soins. La documentation de ces traitements est importante car elle soulève des questions de provenance ou de biais qui doivent être explicitées pour l'interprétation des résultats d'une étude.

Exemple 3 *(Ex. 1 cont.) Reprenons maintenant l'exemple de l'étude sur les cancers du poumon menée sur un entrepôt hospitalier dont on connait le schéma de base de données. Ce que permet la plateforme proposée, c'est de représenter l'étude sous la forme d'un* workflow *computationnel. En pratique, il s'agit simplement d'un fichier (e.g., au format SnakeMake* [6] *) qui décrit l'enchainement des étapes et qui peut être exécutée de manière informatique. Chaque étape fait appel à un code spécifique et paramétrable issu d'une bibliothèque de codes usuels. Par exemple, l'outil de repérage des actes médicaux à base de TAL qui serait paramétrisé dans le* workflow *pour identifier les chirurgie d'exérèse pulmonaire pour sélectionner les patients et patientes d'intérêt. Dans ce cas, le* workflow *conserverait le type et version du modèle TAL utilisé ainsi que l'ensemble des paramètres qui ont été utilisés. De la sorte, le* workflow *décrit complètement le processus et celui-ci peut être (re-)exécuter à la demande de manière identique.*
Ce workflow *aura été construit de manière interactive par l'epidémiologiste ou* data scientist *qui a choisi les traitements élémentaires et les a configurés pour obtenir l'étude souhaitée.*

Certaines approches d'analyse de données médico-administratives s'apparentent à l'idée de concevoir des *workflows* computationnels sous la forme algébrique (*14*) ou logicielle (*15*, *16*). Ces différentes approches apportent une formalisation propre des études épidémiologiques et facilitent l'automatisation des traitements. Elles restent néanmoins difficilement réutilisables par manque d'organisation du partage des ressources.

Un cas applicatif typique serait par exemple la recherche de nouveaux profils pathologiques à partir d'un corpus de données multimodales (textes, images, nombres...) obtenues à partir de dossiers de santé médico-administratifs ou encore des applications de repositionnement de médicaments. Traditionnellement, toutes les étapes décrites précédemment dans la génération d'un *workflow* répondant à ce genre

5. Un modèle d'IA est à comprendre ici comme obtenu par apprentissage automatique. La construction de tels modèles n'est pas l'objectif de notre plateforme : ils sont à construire en dehors et peuvent être utilisés pour enrichir les traitements de données de santé.

6. https ://snakemake.readthedocs.io

d'études auraient été construites en grande partie *ad hoc* et conjointement entre épidémiologistes et personnes expertes en science des données. L'enjeu est de systématiser l'organisation de ses ressources (de données et de *workflows*), tout en favorisant la réflexion éthique et scientifique.

4 De l'outil FAIR au pro-éthique

Concevoir la plateforme selon un principe de *metadesign* – celui d'un bac à sable de développement de *workflows* épidémiologiques – permet d'accommoder ces enjeux organisationnels mais n'est pas suffisant. Suivre un principe d'*open source*, permettant la co-construction de modules intégrables, favorisera l'adéquation de la plateforme aux usages des épidémiologistes. Ceci rejoint pragmatiquement les enjeux d'explicabilité, récurrents dans le cadre de l'éthique de l'IA (*17*), en permettant à l'utilisation de recourir aux méthodes d'explicabilité les plus pertinentes contextuellement. Un partage des *workflows* et résultats obtenus sur la plateforme pourra être réalisé et formalisé différemment en fonction des attentes et des compétences scientifiques et techniques des différentes parties prenantes.

Notre proposition est de développer une plateforme pour concevoir des *workflows* computionnels en épidémiologie en s'appuyant sur des outils du web sémantique (format RDF (*Resource Data Framework*) pour la représentation du schéma des entrepôts, de représentation des méta-données, des workflows et de leurs étapes (*18*, *19*) ; utilisation d'ontologies médicales pour enrichir la sémantique des descriptions) pour s'approcher au mieux des principes FAIR, qui soutiennent la transparence et la reproductibilité du traitement de données. Les *workflows* computationnels scientifiques ont déjà montré leur intérêt pour la FAIRification, notamment dans le cadre de l'analyse de données biologiques (*8*) et ont contribué à l'amélioration de la méthodologie en bioinformatique et à la standardisation de certains outils d'analyse.

L'utilisation d'outils du web sémantique vise ainsi à formaliser les notions de *workflow* en épidémiologie tout en permettant leur documentation par des méta-données facilitant leur réutilisation correcte. Les méthodologies et *worflows* employés par chaque étude seraient transposables aisément vers de nouveaux projets, favorisant la réutilisabilité et la comparaison. Ces propriétés sont ainsi dans la lignée des principes FAIR, promouvant transparence et reproductibilité.

Exemple 4 *Dans l'exemple de notre étude, l'utilisation d'une représentation en RDF permet de décrire des méta-données riches sur le* workflow *de manière standardisée (*via*l'utilisation d'ontologie) et chacune des étapes. Par ailleurs, c'est également toutes les méthodes de la librairie de traitements qui seraient enrichies de méta-données pour les rétrouver et les réutiliser d'une étude à l'autre.*

Les enjeux éthiques liés au développement d'un tel projet ne se limitent toutefois pas à la transparence ni à sa reproductibilité, ni même au caractère co-construit de son développement. Nous reconnaissons la difficulté – voire l'impossibilité de principe – de rendre éthique un processus d'automatisation, même partiel, de l'épidémiologie. L'enjeu devient de rendre le système non pas éthique mais *a minima* pro-éthique (*20*). Cette démarche de conception, proposée par Luciano Floridi, consiste à structurer uniquement l'information fournie lors de procédures décisionnelles, sans que cela n'entrave en rien les options pratiques associées à la décision effective. Le *metadesign* nous paraît donc être parfaitement adapté à ces fins : la plateforme devient avant tout une structure supportant la mise en place de *workflows* épidémiologiques. Dans une démarche pro-éthique, des informations sont données à l'épidémiologiste sur les contraintes statistiques, éthiques ou légales qui peuvent entourer les différentes briques de base mobilisées – données ou raisonnements par exemple. Le but n'est alors plus de faire en sorte que le système garantisse une conformité éthique universelle ou même seulement contextuelle, mais plutôt qu'il fournisse tous les éléments nécessaires à l'épidémiologiste pour construire un *workflow* performant en pleine connaissance des enjeux éthiques sous-jacents. Cela nécessite ainsi de rajouter une couche informationnelle à tous les niveaux de la plateforme. Que ce soit sur la provenance des données – leur représentativité, les biais dans leur collecte, sur les contraintes portant sur certaines briques de bases des modèles à construire – hypothèses de répartition de population par exemple, ou encore sur les résultats mêmes en sortie de pipeline. Adopter une démarche pro-éthique requiert ainsi une objectivation non seulement du processus réflexif épidémiologique (quelles données sont nécessaires, pour quelles fins, au travers de quels traitements), mais également du processus de délibération éthique.

Concevoir une plateforme pour la conception pro-éthique de *workflows* demande ainsi un travail d'anticipation des briques de bases qui devraient pouvoir être intégrées au sein du système. Celui-ci se doit d'être l'intégration technique non seulement des spécifications nécessaires à la réflexion épidémiologique mais il doit également rendre aisément possible la réflexion et l'agir éthique. Cette exigence éthique est récurrente dans la littérature sur le développement responsable de l'IA et plus particulièrement du ML. L'éthique de l'IA est fréquemment fondée sur le suivi de grands principes généraux (*21*). Dans le cadre de projets particuliers, ces derniers s'avèrent souvent abstraits mais peuvent s'incarner sous forme de chartes éthiques fournissant un cadre plus adapté aux contextes d'usage. Si les chartes éthiques ont un rôle majeur à jouer dans la conception technique, cela se doit d'être reflété dans la documentation technique (*22*). De cette façon, tout un écosystème informationnel doit se former autour de l'objet technique, à la fois pour documenter sur ses possibilités, tout autant que pour mettre en garde sur ses potentiels mésusages. En ce sens, le développement de la plateforme de façon pro-éthique devra passer par le développement d'une charte éthique fixant des exigences informationnelles sur chacun de ses éléments. Cela, non dans une visée de transparence pour elle-même mais dans l'objectif de produire une meilleure réflexion épidémiologique dans l'ensemble – éthiquement comme épistémiquement.

Augmenter le niveau d'information sur chacun des éléments de la plateforme n'est pas une réponse satisfaisante en soi. En effet, cela engendrait des problèmes d'interprétabilité communs à tous les systèmes dont l'échelle est trop grande pour être appréhendée cognitivement. L'automatisation partielle peut être possible à ce niveau également. C'est-à-dire en opérant une réflexion sur le *workflow* lui-même. L'utilisation d'une formalisation des *workflows* par des outils du web sémantique rend possible des raisonnements sur les *workflows* et leurs méta-données. L'exemple paradigmatique de ce genre d'implémentation est la quantification de l'équité (*23*) : si chaque traitement indiquait, au travers de ses méta-données, des mesures d'(in-)équité de ces traitements, il serait techniquement possible de quantifier des risques d'inéquité du *workflow* dans son intégralité. Il ne s'agit pas ici de faire en sorte de les optimiser en suivant des standards mathématiques arbitraires mais plutôt d'implémenter de façon native dans la plateforme diverses mesures d'équité permettant d'améliorer la prise en compte de conceptions diverses de l'équité [7] par les systèmes que nous développons (*24*). Il est encore incertain que d'autres principes éthiques puissent se prêter à une formulation sous forme de métrique (*23*). Néanmoins, l'intégration *a minima* de mesures diverses de l'équité pourrait déjà informer l'épidémiologiste tout au long de sa conception pour favoriser certains choix de modules.

Dans cette situation encore, la possibilité de comparer différents *workflows* de traitement de données au sein d'un même cadre permet de faciliter à la fois l'optimisation en fonction de paramètres internes à un *workflow* donné mais également d'établir ensuite une comparaison entre études. L'enjeu est également de permettre d'identifier des moyens d'action en faveur de groupes traditionnellement défavorisés par les études épidémiologiques, tout en permettant une fouille automatisée des données qui pourrait mettre en avant des sous-populations négligées traditionnellement non identifiées.

Au global, une mise en place réussie de la plateforme telle que nous la concevons devrait ainsi permettre à la communauté des épidémiologistes de disposer d'un outil répondant à plusieurs besoins éthiques et scientifiques dans la conception de leurs études. Formalisation des raisonnements (et des hypothèses sous-jacentes), FAIRification et intégration de raisonnements sur la provenance ou encore l'équité, le tout couplé à des alertes concernant d'éventuels défauts dans le raisonnement statistique sont autant d'éléments qui enrichissent les capacités réflexives de l'épidémiologiste.

5 Discussion et perspectives

L'épidémiologie traverse actuellement une période de grandes transformations du fait de la numérisation massive de ses pratiques. Cela entraîne une nécessaire montée en compétences pour les épidémiologistes dans le but de mobiliser du mieux possible les données médico-administratives à disposition. La conception d'une plateforme permettant la construction simplifiée de *workflows* épidémiologiques répond en partie à ces enjeux. La plateforme permet de conceptualiser la réflexion épidémiologique à trois niveaux. Premièrement, elle sert d'interface avec des bases de données existantes au travers d'un formalisme générique de représentation. Ensuite, le traitement de ces données est lui aussi formalisé sous forme d'étapes élémentaires de *workflows*, dont la structure doit être interopérable et documentée. Enfin, le *workflow* lui-même peut devenir l'objet de raisonnements, ouvrant à la possibilité de réflexions épistémologiques et éthiques, de comparaisons inter-*workflows* ou encore d'évaluation de la transformation opérée sur les données de la cohorte initiale. Ceci pourrait faciliter l'évaluation de la robustesse d'une hypothèse entre différents contextes, le croisement de jeux de données, la comparaison de différents *workflows* sur des populations variées et ainsi de suite, tout en associant des indications techniques et éthiques à chaque étude.

Exemple 5 *Dans le cas de l'étude du cancer, nous pouvons illustrer deux raisonnements automatiques qui pourraient être menés :*

- *identification de biais par l'analyse de la provenance : à partir d'un* workflow*, il est possible de remonter jusqu'aux modèles de* machine learning *qui ont été utilisés pour l'étude (par exemple, celui qui a servi à identifier les patients opérés). Admettons que le modèle a été décrit par des meta-données de description du jeu d'apprentissage et qu'elles mentionnent un déséquilibre de genre. Il est alors possible d'inférer (automatiquement) que le* workflow *s'appuie sur des modèles biaisés (et d'informer l'épidémiologiste en conséquent).*
- *explicitation de différences entre études. Si un autre* workflow *a été construit pour répondre à la même question, la mise à disposition des* workflows *permet 1) de répliquer l'autre méthodologie sur ses propres données pour identifier si les différences de résultats viennent des données ou du* workflow *et 2) de comparer formellement les* workflows *(comparaison de graphes) pour identifier ce qui diffère dans la méthodologie.*

Concevoir la démarche de développement d'une plateforme pour l'apprentissage machine en épidémiologique suivant une démarche de *metadesign* pro-éthique se démarque d'une croyance en la possibilité d'intégrer et de garantir l'éthique dans la technique même. La plateformisation de l'épidémiologie telle que nous la concevons ne revient pas à s'assurer de la présence de garde-fous exhaustifs garantissant un caractère éthique sans faille à tous les futurs *workflows* produits *via* la plateforme. Concevoir une plateforme pour rendre éthique la conception d'études épidémiologiques implique donc avant tout pour la plateforme de favoriser la réflexion éthique chez l'épidémiologiste. En ce sens, les enjeux de documentation et de FAIRification se trouvent renforcés par le *design* pro-éthique. Ce dernier s'incarne dans le *metadesign* au travers de la synergie entre

7. Les mesures à prendre pour corriger un problème de parité statistique lié à des données historiquement biaisées ne seront pas les mêmes qu'un biais du traitement impliquant des faux négatifs ou positifs pour un groupe donné.

simplification des outils de construction de *workflows* pour des parties prenantes moins spécialistes et mise à disposition renforcée de l'information. Rendre pro-éthique une plateforme épidémiologique comme nous les concevons ne se substitue ni à une réflexion éthique constante ni à un contrôle extérieur de son caractère éthique. Si elle n'offre pas une garantie absolue d'une épidémiologie éthique, elle fournira néanmoins un terreau fertile à cela. Aucun artifice technique ou critère éthique supposé universel ne peut se substituer à la réflexion éthique et scientifique de l'épidémiologiste peut mener en contexte. L'enjeu est ainsi d'outiller l'épidémiologie des meilleurs outils grâce aux dernières évolutions techniques. Les réflexions philosophiques offertes ici sont préliminaires et ne sont en rien exhaustives. D'autres enjeux philosophiques – paternalisme, asymétries de connaissances, éthique globale... – seront à adresser lors des futurs développements de la plateforme.

Remerciements

La thèse d'É. Pardoux est financée par le projet CNRS Prime-80 ED-AIM.

Références

(*1*) WIENS, J. et SHENOY, E. S. (2017). Machine Learning for Healthcare : On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases 66*, 149-153.

(*2*) SALERNO, J. et al. (2023). Current ethical and social issues in epidemiology. *Annals of Epidemiology 80*, 37-42.

(*3*) DAB, P. W. (1989). L 'ÉPIDÉMIOLOGIE. *G. Brücker et D. Fassin, Santé publique. Aubin Imprimeur : Ligugé, Poitiers*, 11-53.

(*4*) COHEN-BOULAKIA, S. et al. (2017). Scientific workflows for computational reproducibility in the life sciences : Status, challenges and opportunities. *Future Generation Computer Systems 75*, 284-298.

(*5*) WILKINSON, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*.

(*6*) GOLDBERG, M. et al. (2008). Bases de données médico-administratives et épidémiologie : intérêts et limites. *Courrier des statistiques 124*, 59-70.

(*7*) SANGARIYAVANICH, E. et al. (2023). Systematic review of natural language processing for recurrent cancer detection from electronic medical records. *Informatics in Medicine Unlocked*, 101326.

(*8*) COHEN-BOULAKIA, S. et LEMOINE, F. (2024). Workflows for Bioinformatics Data Integration. *Biological Data Integration : Computer and Statistical Approaches*, 53-85.

(*9*) PESSACH, D. et SHMUELI, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR) 55*, 1-44.

(*10*) LIU, B. et al. (2021). When machine learning meets privacy : A survey and outlook. *ACM Computing Surveys (CSUR) 54*, 1-36.

(*11*) FISCHER, G. in *Human-Computer Interaction*, JACKO, J. A. et STEPHANIDIS, C., éd. ; CRC : 2003, p. 88-92.

(*12*) GUENNEC, Y. L. (2016). Le métadesign, ou comment l'expérience doit échapper au designer. *Sciences du Design 4*, 124-127.

(*13*) GOBLE, C. et al. (2020). FAIR Computational Workflows. *Data Intelligence 2*, 108-121.

(*14*) GUYET, T. Enhancing sequential pattern mining with time and reasoning, thèse de doct., Université de Rennes 1, 2020.

(*15*) PETIT-JEAN, T. et al. eds-scikit : data analysis on OMOP databases.

(*16*) BACRY, E. et al. (2020). SCALPEL3 : a scalable open-source library for healthcare claims databases. *International Journal of Medical Informatics*, 104203.

(*17*) NAGAHISARCHOGHAEI, M. et al. (2023). An Empirical Survey on Explainable AI Technologies : Recent Trends, Use-Cases, and Categories from Technical and Application Perspectives. *Electronics 12*, 1092.

(*18*) BOWERS, S. et LUDÄSCHER, B. in *International Workshop on Data Integration in the Life Sciences*, 2004, p. 1-16.

(*19*) SAWADOGO, P., GUYET, T. et AUDUREAU, E. in *Santé et IA 2022*, 2022.

(*20*) FLORIDI, L. (2016). Tolerant Paternalism : Pro-ethical Design as a Resolution of the Dilemma of Toleration. *Science and Engineering Ethics 22*, 1669-1688.

(*21*) JOBIN, A., IENCA, M. et VAYENA, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1*, 389-399.

(*22*) PISTILLI, G. et al. (2023). Stronger Together : on the Articulation of Ethical Charters, Legal Tools, and Technical Documentation in ML.

(*23*) PALUMBO, G., CARNEIRO, D. et ALVES, V. (2024). Objective metrics for ethical AI : a systematic literature review. *International Journal of Data Science and Analytics*.

(*24*) MITTELSTADT, B., WACHTER, S. et RUSSELL, C. The Unfairness of Fair Machine Learning : Levelling down and strict egalitarianism by default, 2023.

Session 4 : Interaction Humain - système d'apprentissage

Expérimentation de la confiance d'un utilisateur de système d'IA

N. Maille[1], K. Amokrane-ferka[2], B. Leblanc[3], N. Heulot[2]

[1] ONERA The French Aerospace Lab, Salon de Provence
[2] IRT SystemX, 91120 Palaiseau
[3] IMS UMR CNRS 5218, ENSC Bordeaux INP, Bordeaux,

kahina.amokrane-ferka@irt-systemx.fr

Résumé

Les systèmes à base d'IA deviennent de plus en plus présents dans la vie de tous les jours et surtout dans les activités professionnelles. Se pose alors la question de savoir ce que les humains éprouvent vis-à-vis de ces systèmes. Leurs attitudes s'apparentent déjà aux relations que peuvent avoir entre eux plusieurs humains, dans le cadre professionnel. Avec l'accroissement des services associés à ces systèmes (aide à la décision, recommandation, etc.), la question de la confiance devient particulièrement critique pour la sécurité et la performance des outils industriels.
L'objectif de ce travail est d'éclairer comment la confiance que l'opérateur se construit dans le système à base d'IA vient modifier son ressenti, son comportement et ses performances globales dans la réalisation de sa tâche. Pour ce faire, un micro-monde a été développé et une étude expérimentale faisant intervenir 32 sujets a été conduite afin de mieux cerner la question. Les résultats montrent une adaptation du comportement en fonction du niveau de fiabilité du système, avec en particulier une confiance rapportée plus forte et une supervision moins importante quand la fiabilité de l'IA est forte.

Mots-clés

Confiance, Système à base d'IA, aide à la décision, collaboration Human-IA.

Abstract

AI-based systems are becoming increasingly present in everyday life, and especially in professional activities. This raises the question of how humans feel about these systems. Their attitudes are already similar to the relationships that several humans may have with each other in the workplace. With the increase in services associated with these systems (decision support, recommendations, etc.), the question of trust becomes particularly critical for the safety and performance of industrial tools.
The aim of this work is to clarify how the trust that the operator builds up in the AI-based system modifies his feelings, behavior and overall performance in carrying out his task. To this end, a micro-world was developed and an experimental study involving 32 subjects was conducted to further investigate the issue. The results show an adaptation of behavior according to the system's level of reliability, with in particular higher reported confidence and less supervision when the AI's reliability is high.

Keywords

Trust, AI based System, decision support system, Human Machine Teaming

1 Introduction

1.1 Contexte applicatif

Notre cadre de travail porte sur un système d'analyse visuelle de conformités, déployé dans des usines automobiles. Ce cas d'usage s'inspire des systèmes conçus pour surveiller différentes parties d'une ligne de production de véhicules, venant prendre des photos de plusieurs points de contrôle spécifiques sous différents angles. Ces points de contrôle peuvent concerner les gaines électriques, les soudures, les peintures, et sont validés visuellement par des opérateurs. Sur la base de ces photos, une IA peut venir analyser la conformité des points et pousser le système à déclencher des alarmes visuelles et sonores à chaque fois qu'une anomalie est détectée (cf Figure 1). L'objectif de ce système est de permettre aux opérateurs, en charge du contrôle qualité, de gagner du temps sur la détection de non conformités afin de disposer de plus de temps pour corriger physiquement ces anomalies lorsque cela est possible.
Ce type de système doit idéalement s'intégrer de manière transparente dans le processus de travail des opérateurs.Or, les premiers retours terrains suite au déploiement de cette solution semblent montrer que les opérateurs sont très sensibles à la robustesse du système. Il semblerait que lorsqu'un système réalise trop souvent des fausses détections ou laisse passer des non conformités, les opérateurs s'en détournent progressivement ce qui interroge les mécanismes de la confiance de l'opérateur dans ce système à base d'IA.

1.2 État de l'art

Les progrès accomplis dans le domaine de l'IA au cours de ces dernières années et leur utilisation croissante dans de nombreux outils, tant de la vie courante que dans l'environnement professionnel, montrent que l'IA va de plus en plus impacter notre quotidien ainsi que nos façons de travailler. Néanmoins, il reste encore des efforts pour ces systèmes passent d'une position de support de la collaboration, à un

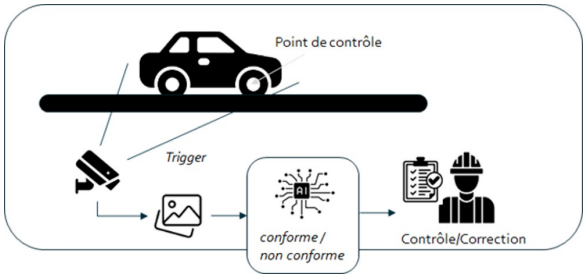

FIGURE 1 – Cas d'usage de conformité visuelle par IA.

système capable d'engagement dans les étapes de la résolution de problème et de prise de décision [12]. La littérature sur le travail d'équipe associant humains et machines identifie la confiance comme un élément essentiel pour que les interactions puissent avoir lieu de manière constructive [2] et ainsi s'inscrire dans une démarche de collaboration.

Cependant, la notion même de confiance, que ce soit dans les individus ou dans les systèmes ne fait pas consensus [13], d'autant que la confiance n'est pas une caractéristique propre au système. Elle s'entend plutôt comme un processus associant un confiant (ou 'trustor') et un mandant (ou 'trustee') [7], couple dans lequel on peut observer l'établissement ou la rupture de ce lien de confiance [11]. Dans la lignée de [3] nous considérons ici que la confiance d'un agent dans un autre représente sa croyance que cet autre agent ne va pas entreprendre des actions qui lui sont préjudiciables. La question de la confiance se pose de manière complexe car cette relation dépend de plusieurs éléments liés au contexte d'interaction [5]. Ces facteurs peuvent être regroupés en trois catégories [3] : ceux liés à l'opérateur, ceux liés à l'automate ou robot et enfin ceux liés à l'environnement. Pour une revue récente de la notion de confiance dans le contexte de la coopération avec des systèmes à base d'IA et des éléments qui l'impactent, le lecteur est invité à se reporter à la méta-analyse de Kaplan et al. [6]. De part sa nature, la confiance n'est pas directement mesurable mais va pouvoir être appréhendée soit à travers le ressenti de la personne faisant confiance (questionnaires), soit par des mesures comportementales car la confiance dans l'autre agent vient aussi modifier leurs interactions [14].

De nombreux facteurs influençant la mise en place de cette confiance ont fait l'objet d'études expérimentales. En particulier, dans le cas où le système d'aide repose sur une IA basée sur de l'apprentissage, l'opacité du système est identifié par la littérature comme nuisant à la confiance qu'il lui accorde l'utilisateur [1]. Ainsi, la manière dont le système explique sa décision peut alors être un facteur déterminant pour l'établissement de ce lien de confiance. Cependant, d'autres facteurs peuvent être tout aussi essentiels. Par exemple l'étude réalisée dans [8], montre l'impact de la performance du système à base d'IA sur la confiance. Les auteurs ont constaté que la confiance dans un système diminue lorsque sa performance est faible. De plus, ils ont souligné que dans les équipes moins performantes, les individus ayant peu confiance dans le système manifestent également peu de confiance envers leurs partenaires humains. Une autre étude [10] s'est focalisée sur l'impact de la nature de l'agent avec lequel un opérateur interagit sur sa prise de décision. L'observation indique que dans une situation où le risque est élevé et où les deux sources d'information (humain et machine) sont équivalentes en termes de fiabilité, le participant préfère choisir une information provenant d'une aide humaine plutôt que celle provenant d'une aide automatisée. La confiance accordée dans l'agent dépendrait donc aussi de sa nature et les résultats acquis pour la confiance entre individus ne peuvent donc pas forcément être étendus à la confiance entre un opérateur et un système à base d'IA. Dans cette même perspective, l'objectif de notre travail est d'éclairer comment la relation de confiance, que l'opérateur se construit avec le système à base d'IA, vient modifier son ressenti, son comportement et finalement ses performances globales dans la réalisation de différentes tâches. Pour cela, le choix est fait de s'appuyer uniquement sur différents niveaux de fiabilité du système d'IA pour modifier cette confiance induite. Dans cet article, nous présentons une expérience de mise en situation d'opérateurs humains dans un micro-monde inspiré du cas d'application de conformité visuelle introduit précédemment.

1.3 Visée de l'étude

Cette étude vise à contribuer à la compréhension de la manière dont un opérateur ressent et s'adapte aux performances d'un système d'assistance à base d'IA. Elle repose sur des sessions de travail répétitives avec ce même système afin que l'opérateur se forge par lui-même une représentation des capacités du système d'aide qui lui est proposé.

Les deux hypothèses formulées sont les suivantes :

- (H1) : l'utilisateur va effectivement percevoir et adapter sa manière de faire en fonction de la fiabilité de l'IA qui lui vient en support (moins l'IA sera fiable, moins l'opérateur lui fera confiance et plus il passera du temps à superviser les résultats proposés par cette IA).
- (H2) : pour une session donnée, le niveau initial de confiance que l'opérateur a dans le système d'aide va impacter sa perception de la fiabilité réelle du système.

Ces hypothèses sont formulées pour répondre aux questions de recherche suivantes :

- Comment la confiance dans le système d'aide modifie-t-elle le comportement de l'utilisateur et ses performances ?
- Comment les performances de l'IA affectent-elles le ressenti de l'utilisateur vis à vis du système ?
- Est-ce que l'on peut conditionner sur la durée, le comportement de l'utilisateur ?

2 Méthodologie

2.1 Participants

Cette étude mobilise un ensemble de 32 participants (15 hommes et 17 femmes) âgés de 18 à 31 ans (M 23.25, ET 3.35), étudiants ou doctorants à l'École Nationale Su-

périeure de Cognitique [1]. Tous ont déclaré avoir une acuité visuelle normale ou corrigée, n'avoir aucun antécédent de troubles neurologiques, et étaient naïfs quant au sujet de l'étude. Ils étaient volontaires pour réaliser l'expérimentation et ont signé un document de consentement éclairé.

2.2 Matériel et stimuli

L'expérimentation s'est déroulée dans un laboratoire de recherche, sur un ordinateur portable (écran 15"). Les sujets étaient installés dans une position proche de la position de travail des opérateurs en atelier (tabouret haut, plan de travail type établis). Chaque passation dure environ 40 minutes. La variable indépendante est la fiabilité de l'IA, et les différentes variables dépendantes mesurées concernent le ressenti de l'opérateur, son comportement et ses performances.

Deux niveaux de fiabilité de l'IA ont été étudiés dans l'expérimentation : (1) une IA très fiable qui ne fait aucune erreur de classification et (2) une IA peu fiable qui fait entre 10% et 30% d'erreur de classification. Chaque sujet travaille avec l'IA pendant 2 blocs consécutifs de 3 essais chacun. Pour un opérateur donné et pour un bloc donné, le niveau de fiabilité de l'IA reste constant ('très fiable' ou 'peu fiable'). Les 32 participants sont répartis sur les quatre combinaisons possibles : 2 (bloc) x 2 (fiabilité de l'IA). Cela constitue donc quatre groupes indépendants de 8 personnes (cf Figure 2).

Groupe	Bloc 1	Bloc 2
G1	Fiable	Fiable
G2	Peu fiable	Peu fiable
G3	Fiable	Peu fiable
G4	Peu fiable	Fiable

FIGURE 2 – Conditions expérimentales associées à chaque groupe.

Les participants ne savent pas à quel groupe ils appartiennent, ils n'ont pas d'information sur la fiabilité de l'IA avec laquelle ils vont travailler et n'ont pas été avertis que cette fiabilité pouvait éventuellement changer d'un bloc à l'autre (donc au bout de 3 essais).

2.3 Procédure

Pour la passation, les participants sont placés dans le contexte d'un travail de classification d'images par distinction des lettres (lettre O) et des chiffres (chiffre 0) à l'aide d'un assistant basé sur un algorithme émulé d'IA (cf Figure 3). Leur tâche consiste à vérifier la validité de la classification proposée par l'IA pour certaines images et de réaliser par eux-mêmes la classification des images que l'IA n'a pas su traiter.

Lors de chaque session, une centaine d'images sont à traiter, comprenant environ 70 images classifiées par l'IA comme

1. https ://ensc.bordeaux-inp.fr

FIGURE 3 – Deux exemples de lettres (à gauche) et de chiffres (à droite).

étant des lettres, environ 10 images classifiées par l'IA comme étant des chiffres et environ 20 images non classifiées. L'interface (cf Figure 4) comprend 3 pages distinctes permettant d'accéder aux images non classées, aux images classées comme des lettres par l'IA et à celles classées comme des chiffres. L'utilisateur peut naviguer librement entre ces trois pages et modifier à son gré la classification de toutes les images, soit pour classer les images non traitées par l'IA, soit pour corriger ce qu'il estime être des erreurs de l'IA. Quand il considère avoir fini sa tâche, le bouton « terminer la session » le conduit à une page lui permettant de donner son ressenti sur la session réalisée, avant de passer à la session suivante. Le temps attribué à chaque session est au maximum de 3 minutes.

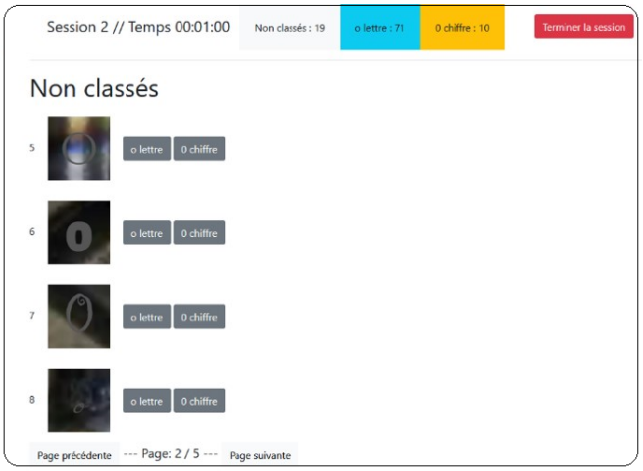

FIGURE 4 – Interface montrant la page des images non classées par l'IA. Les boutons en haut de la page (gris clair, bleu et orange) permettent d'accéder aux trois types de pages. Pour chaque image les deux boutons gris associés permettent de classer l'image dans une catégorie (le bouton devient alors coloré). Les boutons gris (page précédente / page suivante) en bas permettent d'accéder aux autres images de la page choisie.

Le déroulement de la passation comprend (cf Figure 5) : (1) un temps d'accueil (présentation de l'expérimentation au sujet), (2) un temps d'entraînement à la tâche et (3) l'expérimentation en elle-même (la réalisation des 6 sessions).

2.4 Données collectées et statistiques

Les variables dépendantes mesurées sont de 3 types. D'abord des mesures subjectives relatives au ressenti de

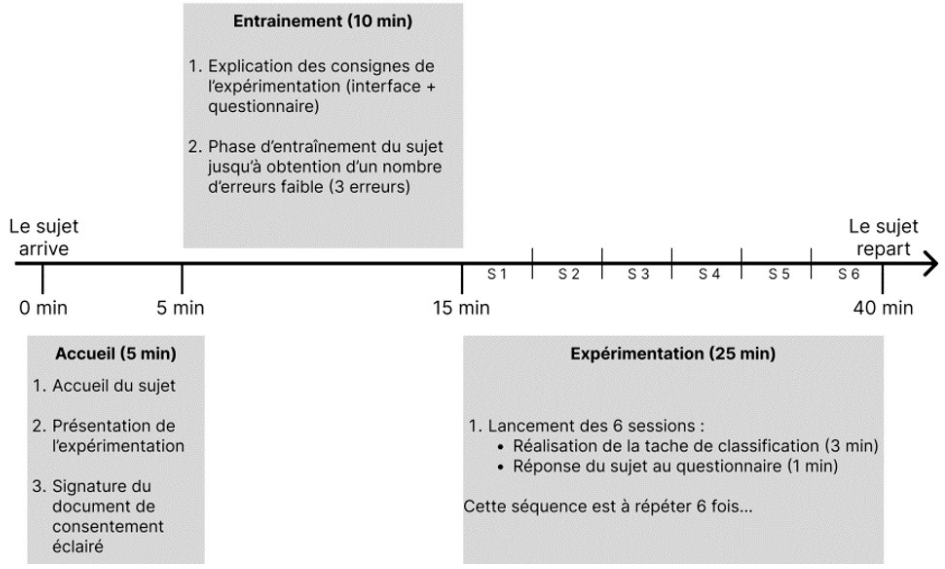

FIGURE 5 – Déroulé d'une passation.

l'opérateur. Elles comprennent : (1) l'évaluation de son niveau de confiance dans la classification réalisée par la machine (VD1 : Confiance_Rapportée), (2) son sentiment de paternité sur le travail fait (i.e. est-ce que la classification vient plutôt de moi ou bien de la machine ; VD2 : Paternité [9]) et (3) son appréciation du taux d'erreur de la machine qui l'amène à demander un ré-entraînement de l'algorithme d'IA (VD3 : ré-entraînement). Pour ces 3 mesures, la réponse est faite sur une échelle non segmentée comprise entre « Pas du tout » et « Totalement ». De plus, une mesure subjective de la charge de travail (VD4 : Charge_Travail) est réalisée à travers une version française du questionnaire du NASA TLX [4] qui comprend 6 dimensions (Exigence mentale, Exigence physique, Exigence temporelle, Succès, Effort, Frustration).

Des mesures comportementales sont recueillies à travers le temps passé sur les différentes pages de l'interface durant l'essai. Le comportement de supervision de l'IA par l'utilisateur est évalué par le temps total passé sur les pages contenant les images classifiées par l'IA (VD5 : Temps_Supervision). Le temps global passé pour réaliser la session est aussi enregistré (VD6 : Temps_Global).

Enfin, deux mesures de performance sont également collectées : le pourcentage d'erreurs de classification faites par le sujet (nombre d'images mal classées par l'opérateur / nombre d'images non classées par l'IA ; VD7 : Erreurs_Classification) ainsi que le nombre d'introduction d'erreurs par le sujet dans la classification de l'IA (image classée correctement par l'IA mais que l'opérateur choisit de classer autrement ; VD8 : Erreurs_Ajoutées).

Des statistiques inférentielles entre les groupes sont réalisées avec un seuil de signification de 0.05. Le terme de 'tendance' est utilisé dans cet article pour un seuil de 0.1. De manière générale, les résultats présentés concernent les comparaisons de 2 groupes indépendants et ces dernières ont été réalisées grâce à des t-tests (test unilatéral).

2.5 Hypothèses détaillées

La première hypothèse de cette étude porte sur la compréhension et l'adaptation de l'opérateur au niveau de fiabilité de l'IA. De manière plus spécifique les hypothèses suivantes sont formulées :

- H1.1 : la confiance dans le système d'aide sera plus faible quand la fiabilité de cette aide est plus faible.
- H1.2 : l'opérateur considérera que sa charge de travail est d'autant plus forte que la fiabilité de l'IA est faible.
- H1.3 : L'opérateur prendra plus de temps pour réaliser la tâche et supervisera plus son système d'aide lorsque celui-ci est peu fiable.
- H1.4 : les performances globales du couple Opérateur/système d'IA seront plus faibles quand l'IA est moins fiable.

La deuxième hypothèse porte sur l'impact de la confiance acquise par l'opérateur envers le système d'IA et sur son évaluation de la fiabilité actuelle de ce système. De manière plus précise, il est attendu que pour une fiabilité du système donnée, une confiance initiale plus basse de l'opérateur l'amène à reporter une confiance moindre dans le système.

3 Résultats

3.1 Hypothèse 1 : Adaptation du comportement

Il est attendu que chaque opérateur évalue (consciemment ou non) pendant les trois sessions du bloc 1 la fiabilité du système d'IA avec lequel il coopère et adapte son comportement à ce niveau de fiabilité. Si la fiabilité de l'IA reste la même, le comportement de l'opérateur pendant les 3 essais du bloc 2 devrait alors être représentatif du comportement final de cet opérateur pour ce niveau de fiabilité de l'IA.

Pour tester si les opérateurs adoptent des stratégies de travail avec l'IA qui dépendent de la fiabilité de l'IA, nous comparons donc les données du bloc 2 pour les groupes G1 et G2. Cette modification du ressenti et du comportement envers l'IA est décomposée en quatre parties que nous évaluons séparément.

H1.1 : fiabilité de l'aide et confiance rapportée

Pour mettre à l'épreuve cette sous-hypothèse, nous utilisons les variables dépendantes VD1, VD2 et VD3. Pour la confiance rapportée (VD1, voir figure 6) il est attendu que celle-ci soit plus faible dans le groupe 2 (IA peu fiable) que dans le groupe 1 (IA fiable). Le test de Student unilatéral montre une différence significative (t(14)=2.99,p<0.005).

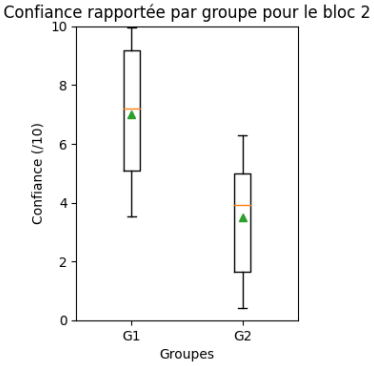

FIGURE 6 – VD1 : Niveau de confiance rapporté par les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

Il est attendu que le sentiment de paternité (VD2, voir figure 7) soit plus fort dans le groupe 2 que dans le groupe 1, ce qui n'est pas le cas (t(14)=-2.59, p=0.01), même si les résultats statistiques indiquent tout juste une tendance.

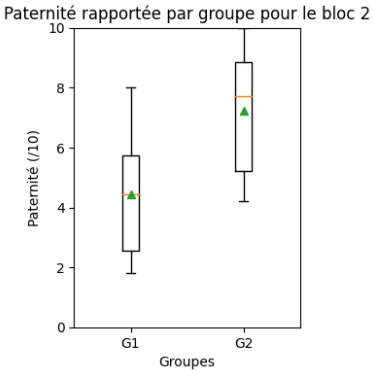

FIGURE 7 – VD2 : Paternité rapportée par les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

Enfin, la demande de ré-entraînement de l'IA (VD3, voir figure 8) devrait être plus forte dans le groupe 2 que dans le groupe 1, ce qui est vérifié (t(14)=-2,75, p<0.001).

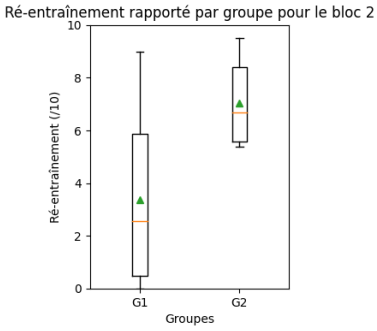

FIGURE 8 – VD3 : Ré-entraînement demandé par les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

H1.2 : fiabilité et charge de travail ressentie

La deuxième composante de cette hypothèse concerne la charge de travail ressentie par l'opérateur qui devrait être plus forte quand l'IA est moins fiable. On attend donc une charge de travail rapportée (VD4, voir figure 9) plus forte dans le groupe 2 que dans le groupe 1, ce qui n'est pas le cas (t(14)=-1.08, p=0.15). En fait seule la composante liée à la pression temporelle augmente de manière significative pour le groupe 2 (t(14)=-1.92, p<0.05).

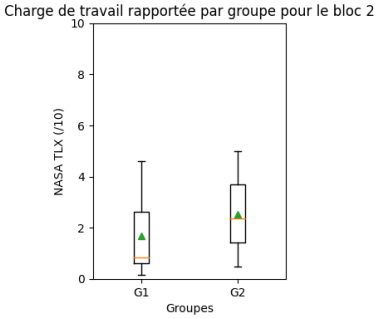

FIGURE 9 – VD4 : Charge de travail rapportée par les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

H1.3 : fiabilité et temps de travail

La troisième partie de l'adaptation de l'opérateur concerne le temps passé par l'opérateur pour réaliser la tâche. Il est attendu que quand la fiabilité de l'IA est plus faible, l'opérateur consacre plus de temps à superviser ce qui lui est proposé (VD5, voir figure 10) et que ceci se répercute sur le temps total passé pour réaliser la tâche (VD6, voir figure 11). Les analyses statistiques confirment une augmentation du temps de supervision de l'IA pour le groupe 2 (t(14)=-2.63, p<0.01) et une augmentation significative du temps total (t(14)=-1.83, p<0.05).

H1.4 : fiabilité et performances

Enfin la dernière partie de cette première hypothèse concerne l'impact de la fiabilité de l'IA sur les performances du couple opérateur-IA pour la réalisation de la

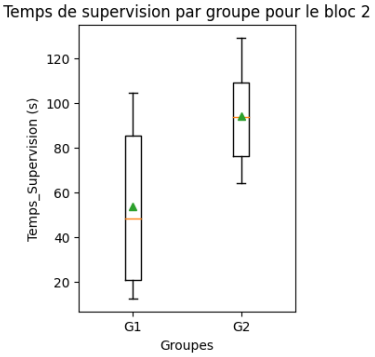

FIGURE 10 – VD5 : Temps de supervision pour les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

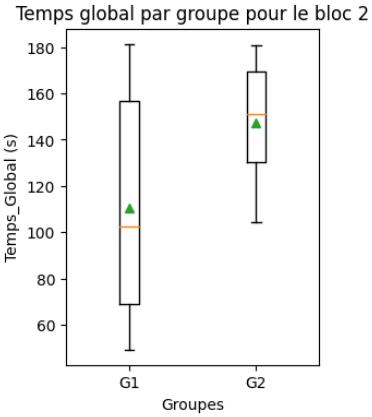

FIGURE 11 – VD6 : Temps total pour réaliser la tâche pour les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

tâche. Il est attendu d'une part que la tâche de classification des images réalisée par l'opérateur seul (VD7) contienne plus d'erreurs quand la fiabilité de l'aide est faible (car l'opérateur est plus occupé par le travail de supervision de l'IA). Le nombre d'erreurs de classification faites par l'opérateur devrait donc être plus important pour le groupe 2, ce qui n'est pas le cas (t(14)=-0.50, p=0.31). De plus il est attendu que l'opérateur introduise plus d'erreurs dans les résultats de classification générés par l'IA (VD8) pour le groupe 2, ce qui n'est pas non plus le cas (t(14)=-0.47, p=0.32).

3.2 Hypothèse 2 : Influence de la confiance acquise

La deuxième hypothèse concerne l'impact de la confiance acquise dans le système d'aide sur la manière dont l'opérateur apprécie les performances de ce système à un moment donné. Pour cela, nous comparons d'une part les groupes 1 et 4 sur le bloc 2 pour regarder si une confiance apprise forte (G1) ou faible (G2) modifie la perception d'une même IA fiable pendant le bloc 2. Puis d'autre part nous regardons les groupes 2 et 3 pour évaluer si une confiance apprise faible (G2) ou forte (G3) impacte l'évaluation d'une même IA peu fiable pendant le bloc 2.

Seules les VD1 (confiance rapportée) et VD3 (nécessité de ré-entraîner l'IA) qui sont les plus caractéristiques du jugement conscient de la qualité de l'IA sont indiquées dans cet article.

Il est attendu ici que la confiance apprise module la confiance rapportée pour la session suivante et que de ce fait la confiance rapportée dans l'IA fiable à la première session du bloc 2 soit plus forte pour le Groupe 1 (confiance apprise forte) que pour le groupe 4. Ceci est confirmé par l'analyse statistique (t(14)=1.91, p<0.05). Il est de plus attendu que la nécessité de ré-entraînement de l'IA (VD3) soit plus élevée pour le groupe 4 que pour le groupe 1 (confiance apprise forte), ce qui n'est pas confirmé par les tests statistiques (t(14)=-1.13, p=0.14).

De même il est attendu que la confiance rapportée (VD1) dans l'IA peu fiable à la première session du bloc 2 soit plus forte pour le Groupe 3 (confiance apprise forte) que pour le groupe 2. Cette conjecture n'est pas confirmée (t(14)=-1.13, p=0.14). Il est de plus attendu que la nécessité de ré-entraînement de l'IA (VD3) soit plus élevée pour le groupe 2 que pour le groupe 3 (confiance apprise forte), ce qui n'est pas non plus confirmé par les tests statistiques (t(14)=-0.46, p=0.67).

4 Discussion

L'expérimentation menée dans cette étude permet de comparer deux à deux les comportements de groupes distincts d'opérateurs qui réalisent une tâche de classification d'images tout en bénéficiant d'un système d'aide à la décision. Le premier bloc de l'expérimentation repose sur trois essais successifs nécessitant de classer à chaque fois une centaine d'images. Ce bloc 1 permet aux sujets de se familiariser avec la tâche mais aussi avec le système d'assistance proposé. Ces essais leur permettent d'optimiser leur manière de coopérer avec cette aide, même si aucune information ne leur est donnée, ni sur la fiabilité de l'aide qui leur est proposée, ni sur la qualité du travail final qu'ils ont réalisé conjointement.

Les deux premiers groupes d'utilisateurs considérés réalisent la tâche soit avec une assistance fiable qui ne fait pas d'erreur de classification des images (Groupe1), soit avec une assistance peu fiable ayant un taux d'erreur de l'ordre de 20%. Les résultats obtenus montrent que les deux groupes d'opérateurs perçoivent cette différence dans la performance du système d'aide et modulent leur comportement en fonction de ce ressenti. Le report conscient de la qualité du système d'aide qui leur est proposé se retrouve tout d'abord dans la confiance attribuée au système. Ceci montre bien que les opérateurs ont supervisé le travail du système d'aide et qu'ils sont sensibles aux erreurs commises par cette IA. La présence d'erreurs dégrade la confiance attribuée à l'aide. En parallèle, et de manière cohérente, les opérateurs collaborant avec le système le moins fiable demandent plus que les autres un réajustement de

cette aide, confirmant cette prise en compte du taux d'erreur de l'aide. La fiabilité de l'aide proposée est donc bien dans cette étude un facteur qui module la perception qu'en a l'opérateur.

Par contre, l'opérateur ne se considère pas fortement plus impliqué dans la classification globale des images. Même si une tendance vers une implication plus forte se dessine, les opérateurs ne s'attribuent pas réellement une part plus importante du travail réalisé. Ceci est corroboré par le fait que la charge de travail ressentie n'est pas différente pour les deux groupes, ce qui tend à montrer que l'opérateur continue à tirer profit de cette aide.

Il y a donc à la fois une acceptation de l'aide et une adaptation de la confiance qui lui est attribuée. Ceci amène l'opérateur à modifier la manière dont il supervise les propositions de cette IA. Les résultats montrent qu'en effet le groupe utilisant une IA moins fiable consacre plus de temps à la superviser, c'est-à-dire à vérifier la validité de la classification des images qu'elle propose. Cette augmentation du temps nécessaire pour valider les propositions du système d'aide ne se fait pas au détriment du temps consacré à sa propre tâche de classification des erreurs, mais en augmentant le temps total utile pour finaliser l'essai.

Cette adaptation comportementale résulte en partie de ce que le temps laissé à l'opérateur pour réaliser la tâche (3 minutes) n'amenait pas une contrainte temporelle forte. Quand la fiabilité de l'aide est faible et que sa supervision devenait plus importante cette limite temporelle commençait à devenir une contrainte plus forte, ce qui est confirmé par le résultat sur la dimension « pression temporelle » de la charge de travail qui est significativement plus forte pour le groupe travaillant avec l'IA de plus faible fiabilité.

Enfin les résultats montrent que, contrairement à ce qui était attendu, cette diminution de la fiabilité de l'aide ne fait pas baisser les performances de l'opérateur dans sa tâche spécifique, ni la performance globale. Il semble donc ici que l'ajustement de la supervision du travail réalisé par le système d'aide ne vienne pas contraindre de manière sensible la tâche de l'opérateur qui garde le même niveau d'exigence pour sa partie de la classification et s'attribue suffisamment de temps pour juger de manière adéquate le travail de son système d'aide.

Ces résultats montrent que quand la pression temporelle n'est pas trop forte, l'opérateur reste investi dans son travail collaboratif avec une IA dont la fiabilité est plus faible. Dans un cadre industriel, par exemple pour un contrôle qualité où le temps disponible pour réaliser l'ensemble des contrôles est souvent assez contraint, la tolérance de l'opérateur à un taux d'erreur plus important du système d'aide pourrait être plus faible. Le protocole présenté dans cette étude reste en mesure d'investiguer ce type de problématique, soit en réduisant le temps disponible pour la tâche, soit en augmentant le taux d'erreur de l'IA jusqu'à observer un point de rupture dans l'utilisation de l'aide. En particulier il serait pertinent de regarder de manière plus précise le comportement de l'opérateur, dans un premier temps avec les stratégies de parcours des différentes pages de l'aide (existe-t-il des stratégies de parcours différentes selon la pression temporelle ou le taux d'erreur ?), et de manière plus fine à l'aide d'un oculomètre (le regarde parcourt-il toutes les images ?).

La deuxième partie des résultats a permis d'explorer si l'apprentissage de la confiance dans le système influençait, à un instant donné, l'évaluation que se fait l'opérateur de la fiabilité du système d'aide. Les résultats obtenus sont moins tranchés. L'évaluation d'une IA fiable semble dépendre de la confiance apprise, le groupe ayant appris à avoir peu confiance restant plus méfiant que l'autre groupe. Par contre, en sens inverse, quand les deux groupes sont confrontés à une IA peu fiable, il ne semble pas y avoir de différence. Ceci tendrait à montrer que, dans ce contexte, il est plus compliqué (plus long) de gagner la confiance que de la perdre. Ce résultat est certainement lié au fait que la situation expérimentale amenait les utilisateurs à rester impliqués dans leur tâche, de par la partie active de classification qu'ils avaient à réaliser et par le fait que la durée était relativement restreinte (6 sessions). Il semblerait que dans ce cadre il n'y ait pas de phénomène de surconfiance qui se soit installé chez les opérateurs, ce qui ne serait pas forcément le cas avec un usage journalier, beaucoup plus répétitif. Le protocole proposé pourrait permettre d'investiguer plus en profondeur cette problématique et chercher s'il est possible de mettre en évidence une habituation au système qui atténuerait les facultés de l'opérateur à identifier une baisse de performance du système.

5 Conclusion

Le travail présenté dans ce papier a consisté à explorer les attitudes d'utilisateurs mis face à un système à base d'IA. Aujourd'hui, expérimentales, ces situations ne vont pas tarder à se généraliser dans le monde professionnel. Le cas d'usage que nous avons choisi s'inspire de situations réelles du contrôle en continu de la qualité d'une production industrielle. Certains de ces contrôles, essentiellement visuels, vont très vite être opérés par des systèmes à base d'IA, charge à l'opérateur de contrôler ce contrôle. Dans ces situations hybrides où humains et machines vont devoir collaborer, la question de la confiance de l'opérateur dans le système à base d'IA devient une question centrale. Pour étudier cette question, nous proposons un dispositif où l'opérateur doit distinguer visuellement, sur une image bruitée, des chiffres zéros et des lettres O. Il s'agit d'une tâche de classification très facilement compréhensible, facile à opérer, mais pouvant demander un fort investissement cognitif en fonction du rythme imposé et du bruit dans les images à classer. L'apport d'un système d'IA y est donc salutaire tout en restant expérimentalement contrôlable dans ses performances.

Nous retrouvons dans ce dispositif expérimental les tendances attendues dans la construction de la confiance que l'opérateur se fait dans le système. Cela s'observe dans les modifications de son ressenti, de ses performances et de son comportement. Les résultats montrent une adaptation du comportement en fonction du niveau de fiabilité du système. C'est ainsi que nous mettons en évidence que

la confiance rapportée est plus forte et la supervision est moins importante lorsque la fiabilité de l'IA est forte. Ce genre d'investigation est nécessaire pour préparer la mise en service d'outils d'aide à la décision dans la production industrielle.

Bien que les premiers résultats soient assez attendus, comme par exemple que la fiabilité de l'IA ait un impact direct sur la confiance qui lui est accordée et sur la manière de la superviser, cette étude montre que ces résultats sont vérifiables empiriquement et apporte un protocole expérimental réutilisable. Ceci ouvre de nouvelles perspectives pour l'étude des interactions entre un opérateur et un système d'IA. Plusieurs pistes sont actuellement explorées.

La première piste porte sur la compréhension des distinctions engendrées par l'utilisation d'un système d'assistance traditionnel par rapport à un système d'aide à base d'IA. Une étude récente (Munoz et al., en cours de soumission) met en lumière l'impact de l'agent proposant une assistance (humain, système automatisé classique ou IA) sur la décision d'un opérateur lorsqu'il doit choisir entre différentes formes d'aide. Les résultats suggèrent que la nature de l'agent influence l'acceptation de l'aide, avec des implications différentes selon qu'il s'agisse d'un système automatisé classique ou basé sur l'IA. Par conséquent, la confiance accordée à ces systèmes et leur acceptation varient en fonction de leur spécificité, ce qui soulève la question de leur utilisation dans des contextes de travail répétitif et de la manière dont la confiance se développe au fil du temps.

La seconde perspective de recherche aborde la question de l'explicabilité des systèmes d'IA et de son influence sur la construction de la confiance envers le système d'assistance. De nombreuses études se penchent sur la nature de ces explications et leur impact sur la collaboration avec l'opérateur, notamment en ce qui concerne l'acceptabilité de l'aide, la confiance dans les décisions prises, et les performances dans l'exécution des tâches (voir par exemple Larasati et al., 2023 ; Le Guillou et al., 2023). Le protocole expérimental proposé dans cette étude offre une opportunité de manière plus pragmatique pour évaluer l'effet de différents types d'explications sur la construction de la confiance dans les systèmes d'IA. Dans cette optique, la fiabilité de la classification demeurerait constante, mais des éléments explicatifs de la classification des images seraient ajoutés dans une des conditions afin de mesurer leur effet sur l'exécution de la tâche (voir par exemple Van der Waa et al., 2021 pour une introduction à l'évaluation des explications dans les systèmes d'assistance basés sur l'IA).

Les recherches menées par Merritt et ses collègues (2015) examinent les représentations que les individus ont des systèmes d'aide et leur impact sur la confiance qu'ils accordent à ces systèmes. Ils démontrent notamment que certaines personnes ont des attentes dichotomiques (confiance totale ou défiance) et que la relation entre la fiabilité de l'assistance et la confiance reportée n'est pas forcement linéaire. De manière plus fondamentale, notre étude vise à éclairer les mécanismes et les facteurs qui contribuent à l'établissement d'un lien de confiance entre un opérateur et un système d'aide, quelle que soit sa nature. Une perspective à explorer ici consiste à examiner expérimentalement la nature de cette relation entre la fiabilité de l'assistance et la confiance établie, en la faisant varier de manière plus graduelle (en testant des niveaux de fiabilité intermédiaires). Cela permettrait d'obtenir des données pour mieux comprendre la transition entre les comportements de confiance et de méfiance envers un système d'aide.

Remerciements

Ce travail a obtenu le soutien du gouvernement français dans le cadre du programme "France 2030", au sein de l'Institut de Recherche Technologique SystemX et du projet Confiance.ai [2].

Références

[1] Plamen P. Angelov, Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. Explainable artificial intelligence : an analytical review. *WIREs Data Mining and Knowledge Discovery*, 11(5), September 2021.

[2] Laurent Chaudron, Jean-Marie Burkhardt, Lisa Chouchane, Pauline Muñoz, Nicolas Maille, and Anne-Lise Marchand. Trust : The vital fluid of interactions. In *AHFE 2023*, 2023.

[3] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5) :517–527, 2011.

[4] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index) : Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[5] Kevin Anthony Hoff and Masooda Bashir. Trust in automation : Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3) :407–434, 2015.

[6] Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. Trust in Artificial Intelligence : Meta-Analytic Findings. *Human Factors : The Journal of the Human Factors and Ergonomics Society*, 65(2) :337–359, March 2023.

[7] Laurent Karsenty. Comment appréhender la confiance au travail. *La confiance au travail*, pages 13–51, 2013.

[8] Nathan McNeese, Mustafa Demir, Erin Chiou, Nancy Cooke, and Giovanni Yanikian. Understanding the role of trust in human-autonomy teaming. 2019.

[9] Adrien Metge. *Opérateurs et systèmes intelligents : se comprendre pour décider. Application à la supervision de drones.* PhD thesis, Université de Bordeaux, 2022.

2. https ://www.confiance.ai/

[10] Pauline Munoz, Anne-Lise Marchand, Laurent Chaudron, and Nicolas Maille. Cl : Confiance en l'autre : approche expérimentale de l'arbitrage entre le partenaire humain et le partenaire automatisé. In *12ème Colloque de Psychologie Ergonomique EPIQUE 2023*, 2023.

[11] Yiteng Pan, Fazhi He, Haiping Yu, and Haoran Li. Learning adaptive trust strength with user roles of truster and trustee for trust-aware recommender systems. *Applied Intelligence*, 50 :314–327, 2020.

[12] Isabella Seeber, Eva Bittner, Robert O Briggs, Triparna De Vreede, Gert-Jan De Vreede, Aaron Elkins, Ronald Maier, Alexander B Merz, Sarah Oeste-Reiß, Nils Randrup, et al. Machines as teammates : A research agenda on ai in team collaboration. *Information & management*, 57(2) :103174, 2020.

[13] Thomas B Sheridan. Individual differences in attributes of trust in automation : measurement and application to system design. *Frontiers in Psychology*, 10 :1117, 2019.

[14] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to Evaluate Trust in AI-Assisted Decision Making ? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2) :1–39, October 2021.

Prédiction de profils étudiants sur une plateforme d'apprentissage en ligne

Pauline Chiquet[1], François Lecellier[1], Philippe Carré[1]

[1] Université de Poitiers, Univ. Limoges, CNRS, XLIM, Poitiers, France

{pauline.chiquet, francois.lecellier, philippe.carre}@univ-poitiers.fr

Résumé

Depuis une vingtaine d'années, les universités utilisent des outils numériques qui génèrent des traces numériques pour améliorer l'accessibilité de leurs cours. L'étude présentée dans cet article a pour objectif de caractériser des profils étudiants de manière automatique à partir de ces données. Cependant, en général, les études dans ce cadre sont confrontées à des classes déséquilibrées. Nous proposons d'analyser l'influence du sur-échantillonnage d'une base de données issue d'une plateforme d'apprentissage en ligne du supérieur de Poitiers. Nos résultats montrent que le sur-échantillonnage permet d'améliorer la précision et le rappel des modèles de prédiction et, par conséquent, de mieux détecter notamment les situations d'abandon.

Mots-clés

Analyse de l'apprentissage, apprentissage supervisé, Moodle, sur-échantillonnage.

Abstract

For the past twenty years, universities have been using digital tools that generate digital traces to improve the accessibility of their courses. The study presented in this article aims to characterize learner profiles automatically from this data. However, in general, studies in this context are confronted with unbalanced classes. We propose to analyze the influence of oversampling on a database from an e-learning platform at Poitiers University. Our results show that oversampling improves the precision and recall of prediction models, and consequently enables better detection of dropout situations in particular.

Keywords

Learning analytics, Moodle, Oversampling, Supervised learning.

1 Introduction

Les Learning Analytics, également appelés analyse de l'apprentissage, consistent à utiliser et analyser des données liées à l'apprentissage et à l'éducation. Les objectifs sont multiples : le suivi de l'acquisition des connaissances, la prédiction de profils ou de résultats, ou encore, la personnalisation de l'enseignement. Les données utilisées sont les traces numériques, les évaluations, les informations démographiques, mais également l'historique académique [1]. Dans cette étude, nous utilisons Motive. Cette plateforme d'autoformation, dédiée aux compétences transversales est construite sous Moodle par l'Université de Poitiers dans le cadre du projet Elans [1]. Elle est composée de 11 chapitres avec 26 tests et 74 modules au total. La figure 1 présente l'organisation de la plateforme Motive. Les chapitres portent sur différentes thématiques telles que la prise de note, la recherche documentaire, la gestion de projet, les compétences numériques, etc.

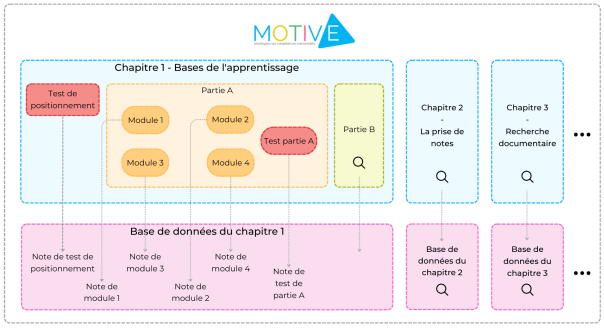

FIGURE 1 – Organisation de la plateforme Motive (zoom sur le chapitre 1). Chaque note obtenue est stockée dans une base de données associée au chapitre.

Chaque chapitre débute par un test de positionnement. Ce test permet de connaître le niveau de l'étudiant. Puis, l'étudiant réalise différents modules à l'issue desquels il obtient des notes de module. Les modules peuvent être réalisés dans n'importe quel ordre. Après avoir effectué tous les modules, l'étudiant réalise un test de partie. Une note de test de partie est alors obtenue. En d'autres termes, les modules correspondent à des exercices d'entraînement portant chacun sur une notion particulière. Les tests de partie correspondent à des évaluations reprenant l'ensemble des notions présentées dans une partie.
Pour chaque chapitre, trois types de fichier sont disponibles : les fichiers de **modules**, de **tests** et de **logs**. Ces derniers contiennent les traces numériques générées par les étudiants sur la plateforme. À partir de ces données, chaque

1. L'Université de Poitiers bénéficie d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme des Nouveaux Cursus Universitaire (NCU ELANS - réf. ANR-18-NCUN-0026).

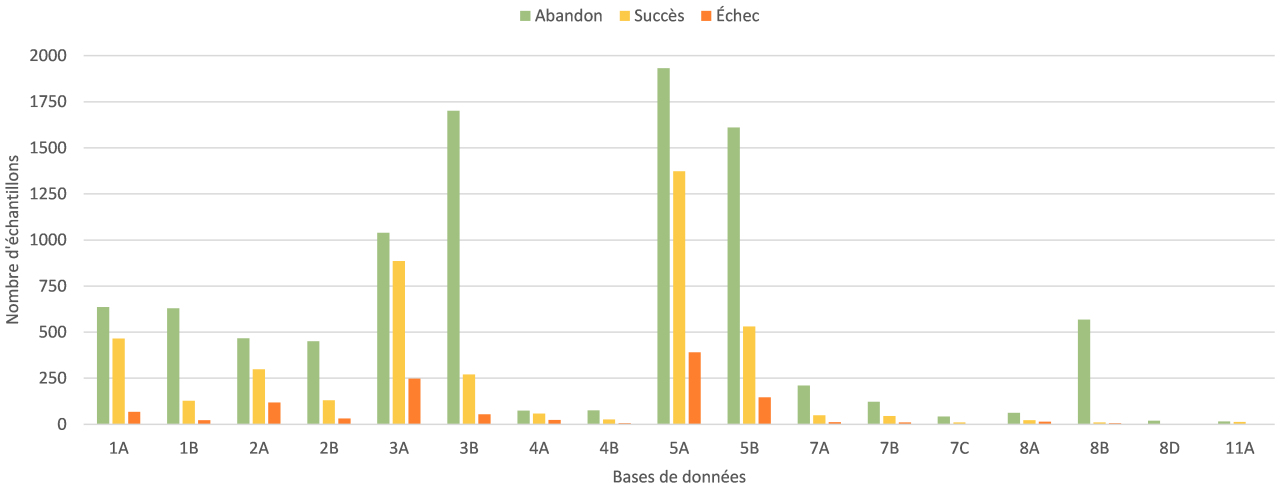

FIGURE 2 – Répartition des profils étudiants au sein des chapitres de Motive

utilisateur a pu être étiqueté, de manière automatique, selon trois catégories : Abandon, Succès ou Échec.
Cet étiquetage permet de connaître la répartition des profils étudiants au sein de Motive (figure 2). Ces données sont très déséquilibrées : la classe Abandon est beaucoup plus représentée que les classes Succès et Échec. Afin de pouvoir travailler sur un ensemble de données statistiquement satisfaisant, nous avons décidé de nous concentrer sur les données issues des parties 2A, 3A, 5A et 5B qui contiennent le plus de données. Il est important de noter que la répartition des classes finales est également fortement déséquilibrée (figure 2).

L'objectif de ces travaux est de prédire les profils étudiants à l'aide d'un modèle de prédiction. Ce modèle permettra de détecter les étudiants en difficulté avant la fin d'un chapitre. Pour cela, nous devons lutter contre le déséquilibre des données. Premièrement, nous présenterons plusieurs méthodes d'analyse de l'apprentissage, mais également des méthodes de sur-échantillonnage des données. Puis, nous expliquerons la démarche de mise en place des modèles de prédiction. Enfin, nous détaillerons l'ensemble des résultats obtenus.

2 État de l'art

2.1 Méthodes de l'analyse de l'apprentissage

L'analyse de l'apprentissage repose sur plusieurs méthodes. Les modèles statistiques sont les plus utilisés (45% des publications selon l'étude de NAMOUN et ALSHANQITI [15]). FOUNG et CHEN [5] se basent sur un modèle de régression pour comprendre comment les étudiants utilisent une plateforme d'apprentissage en ligne, mais également pour prédire la notion de succès en fonction des traces numériques. Afin de pallier la difficulté de prédiction, les auteurs suggèrent de combiner les données liées aux interactions avec la plateforme d'apprentissage à des données externes (par exemple, des informations géographiques ou des antécédents scolaires). De leur côté, MING et MING [13] utilisent l'analyse sémantique latente probabiliste (PLSA) et l'allocation de Dirichlet latente (LDA) pour prédire les notes finales des étudiants à partir des forums de discussion en ligne. Les résultats obtenus par cette combinaison d'approche sont prometteurs.
Les méthodes de Machine Learning sont également très utilisées pour comprendre les profils d'apprentissage des étudiants et les prédire. MORENO-MARCOS et al. [14] ont analysé plusieurs facteurs afin de connaître leur influence sur la prédiction de la performance d'un étudiant. Le jeu de données contient les interactions avec les exercices et les forums, la liste des vidéos qui ont été ouvertes et les suivis des clics. Deux prédictions sont possibles : succès ou échec. Pour analyser l'influence des facteurs, les auteurs utilisent quatre types d'algorithmes standards d'apprentissage supervisé. Les résultats montrent que les données liées aux exercices sont de très bons indices de prédiction. À contrario, les données liées aux forums ou aux clics n'ont généralement que peu d'impacts.
Certains auteurs se sont intéressés au problème de l'apprentissage non supervisé. KUZILEK et al. [11] ont vérifié la corrélation entre les interactions des étudiants avec une plateforme d'apprentissage en ligne et les notes obtenues. Le jeu de données utilisé est OULAD [10]. Les chercheurs utilisent l'algorithme d'espérance-maximisation pour regrouper les données d'interactions des étudiants selon six classes. Les résultats montrent que certains de ces groupes présentent des performances élevées, tandis que d'autres montrent des signes de difficultés dès le début du cours.
FRANCIS et BABU [6] décrivent un modèle de prédiction des résultats des étudiants à partir de données issues de l'enseignement supérieur de l'État Kerala en Inde. Ces données sont organisées selon quatre types de caractéristiques telles que des caractéristiques démographiques, académiques, liées aux interactions avec la plateforme et supplémentaires. Trois prédictions sont possibles : bon résultats, résultats moyens et résultats faibles. La fouille de

données est utilisée via quatre algorithmes de classification standards afin de déterminer les caractéristiques ayant le plus d'impact sur les résultats. Puis, ces caractéristiques sont utilisées comme entrée pour l'algorithme de clustering (K-moyennes [12]). Les résultats montrent une forte corrélation entre les interactions de l'étudiant avec la plateforme d'apprentissage en ligne et ses résultats académiques. Ce type de modèle permet d'obtenir une précision de 0,75.
Suivant le même principe, certains chercheurs combinent l'apprentissage supervisé et l'apprentissage non supervisé. C'est le cas de IATRELLIS et al. [9] qui présentent une méthode permettant de prédire les résultats des étudiants de licence afin de savoir s'ils pourront poursuivre, ou non, leurs études. Les données incluent la moyenne des notes finales, le parcours suivi, des notes de projet, le nombre de redoublement, le rang au sein de la promotion, etc. La première étape consiste à regrouper les étudiants ayant des données similaires à l'aide d'un algorithme de K-moyennes [12] (partie non supervisée). Finalement, trois groupes ont été retenus ($k = 3$). Puis, une forêt d'arbres décisionnels [3] (partie supervisée) est utilisée pour prédire si les étudiants de licence pourront poursuivre leurs études en master ou non. Les résultats montrent qu'un modèle de prédiction précédé d'un clustering obtient de meilleures performances qu'un modèle de prédiction seul.

2.2 Méthodes de sur-échantillonnage

L'ensemble des méthodes d'analyse de l'apprentissage nécessitent une quantité importante de données pour éviter tout risque de sous-apprentissage des modèles. Les données doivent également être suffisamment équilibrées afin de ne pas introduire de biais lors de l'apprentissage. WONGVORACHAN, HE et BULUT [18] présentent trois méthodes pour corriger ce type de déséquilibre : le sur-échantillonnage, le sous-échantillonnage et l'échantillonnage hybride. Le sur-échantillonnage consiste à augmenter le nombre d'échantillons de la ou des classes minoritaires. À l'inverse, le sous-échantillonnage consiste à diminuer le nombre d'échantillons de la ou des classes majoritaires. Enfin, l'échantillonnage hybride consiste à diminuer le nombre d'échantillons de la ou des classes majoritaires et à augmenter le nombre d'échantillons de la ou des classes minoritaires. Du fait du déséquilibre important de nos données, nous avons décidé de nous intéresser plus particulièrement au sur-échantillonnage. La méthode la plus simple est le sur-échantillonnage aléatoire (ou ROS pour Random Oversampling). Celle-ci consiste à dupliquer aléatoirement des échantillons de la classe minoritaire avec remplacement jusqu'à ce que la proportion des deux classes soit équilibrée [18]. La figure 3 illustre cette méthode. Cette méthode est très simple mais occasionne un risque de sur-apprentissage [4].
Il est également possible de créer des échantillons synthétiques à partir des échantillons originaux. Selon CHAWLA et al. [4], le sur-échantillonnage synthétique de la classe minoritaire (ou SMOTE pour Synthetic Minority Oversampling Technique) consiste à augmenter le nombre d'échantillons de la classe minoritaire par la création d'échantillons synthétiques. La classe minoritaire est sur-échantillonnée selon l'algorithme simplifié suivant :

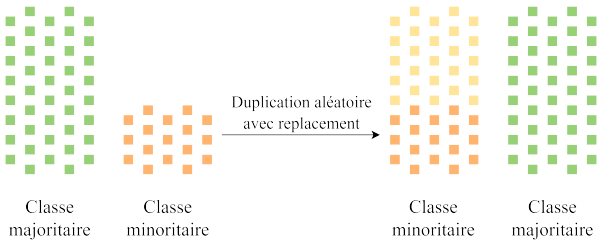

FIGURE 3 – Méthode du sur-échantillonnage aléatoire. En vert, la classe majoritaire. En orange, la classe minoritaire.

1. Sélection aléatoire d'un échantillon de la classe minoritaire ;
2. Identification des k plus proches voisins parmi la classe minoritaire ;
3. Création d'échantillons entre l'échantillon sélectionné et ses k plus proches voisins (par interpolation sur les caractéristiques des échantillons).

Ces étapes sont répétées jusqu'à ce que le nombre d'échantillons de la classe minoritaire soit équivalent à celui de la classe majoritaire. La figure 4 schématise cette méthode.

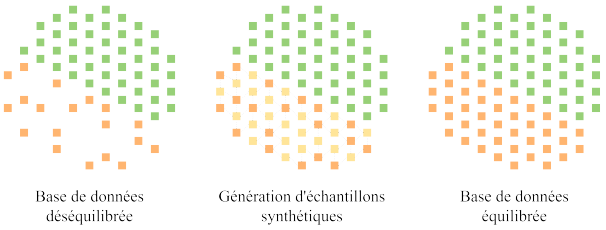

FIGURE 4 – Méthode SMOTE. En vert, la classe majoritaire. En orange, la classe minoritaire. En jaune, les échantillons synthétiques appartenant à la classe minoritaire.

Ces deux méthodes de sur-échantillonnage ont été utilisés par les auteurs HASSAN, AHMAD et ANUAR [8]. L'objectif était de prédire des profils étudiants à partir de données démographiques, académiques et de journaux d'activités. Les profils ont été établis en se basant sur une moyenne pondérée des notes obtenues, classées ensuite en trois catégories : faible, moyenne et excellente. Les données ont été rééquilibrées à l'aide de sur-échantillonnage, de sous-échantillonnage et d'échantillonnage hybride. Les modèles (forêts d'arbres décisionnels) issus du sur-échantillonnage obtiennent un F-score de 0,870 pour le ROS et 0,750 pour le SMOTE. Le meilleur F-score est obtenu avec l'algorithme AdaBoost et le sur-échantillonnage ROS. Les autres algorithmes obtiennent des F-scores inférieurs.
Enfin, nous pouvons également citer deux méthodes de sur-échantillonnage issues du SMOTE. Selon HAN, WANG et MAO [7], la méthode de Borderline-SMOTE consiste à sur-échantillonner uniquement les échantillons minoritaires li-

mites. En d'autres termes, cette méthode génère des échantillons synthétiques de la classe minoritaire uniquement près de la frontière avec la classe majoritaire. La figure 5 illustre cette méthode.

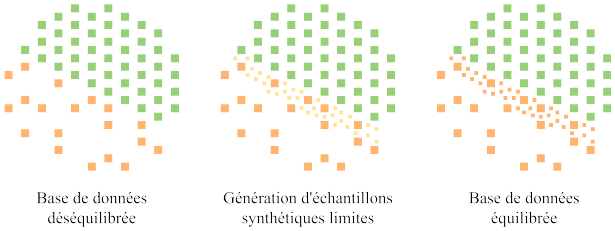

FIGURE 5 – Méthode Borderline-SMOTE. En vert, la classe majoritaire. En orange, la classe minoritaire. En jaune, les échantillons synthétiques limites appartenant à la classe minoritaire.

Il existe également le SVM SMOTE. Il s'agit de la combinaison de la méthode SMOTE et de l'algorithme des machines à vecteurs de support (SVM). Selon NGUYEN, COOPER et KAMEI [16], cette méthode consiste à déterminer les limites des classes à l'aide d'un SVM, puis à générer des échantillons synthétiques de la classe minoritaire. Ces deux dernières méthodes sont intéressantes dans les cas où les échantillons à classifier sont relativement ressemblant. Elles permettent de renforcer les caractéristiques des limites des classes.

3 Modèles de prédiction

3.1 Attributs des nouvelles bases de données

Les données issues de Motive ne sont pas directement utilisables par des modèles de prédiction. Ainsi, des nouvelles bases de données ont été construites à partir des données issues de Motive. Chaque chapitre possède les attributs suivants :

— **Identifiant de l'utilisateur** : `userid` commun aux trois types de fichiers ;

— **Modules** : nombre de tentatives de chaque module pour chaque utilisateur ;

— **Temps moyens des modules** :
  - Temps moyen module x : temps moyen pour un module ;
  - Temps moyen total : temps moyen pour l'ensemble des modules réalisés (si un ou plusieurs modules ne sont pas réalisés, ils ne sont pas pris en compte dans le calcul du temps moyen total).

— **Notes moyennes des modules** :
  - Note moyenne module x : note moyenne pour un module ;
  - Note moyenne totale : note moyenne pour l'ensemble des modules réalisés (si un ou plusieurs modules ne sont pas réalisés, ils ne sont pas pris en compte dans le calcul de la note moyenne totale).

— **Test de positionnement** : indique si l'étudiant a réalisé le test de positionnement ou non ;

— **Note du test de positionnement** : note obtenue par l'étudiant au test de positionnement ;

— **Logs** : agrégation des données de logs pour les 101 événements ;

— **Classe de l'échantillon** : Trois classes sont possibles :
  - **Abandon** : l'étudiant a commencé le module mais n'a pas réalisé ou n'a pas terminé le test de partie ;
  - **Succès** : l'étudiant a commencé le module et a terminé le test de partie avec une note supérieure ou égale à 80% de la note maximale ;
  - **Échec** : l'étudiant a commencé le module et a terminé le test de partie avec une note inférieure à 80% de la note maximale.

3.2 Description des pipelines

L'objectif est de prédire le profil des étudiants malgré une base de données déséquilibrée. Trois pipelines seront testés pour déterminer l'impact du sur-échantillonnage sur ces données.

— **Pipeline 1** : Pas de sur-échantillonnage (figure 6).

— **Pipeline 2** : Sur-échantillonnage avant la séparation des données (figure 7).

— **Pipeline 3** : Sur-échantillonnage après la séparation des données (figure 8).

Premièrement, les données sont normalisées à l'aide de la formule suivante :

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

où z est la valeur centrée réduite, x est la valeur à normaliser, μ est la moyenne de toutes les valeurs à normaliser et σ est l'écart-type de toutes les valeurs à normaliser.
Après cette phase de normalisation, nous séparons les données selon une pondération 70/30 en prenant soin de maintenir la répartition des classes dans chacun des deux échantillons.
Puis, les données sont sur-échantillonnées avant (pipeline 2) ou après (pipeline 3) la séparation des données. Deux méthodes de sur-échantillonnage sont testées : le sur-échantillonnage aléatoire (ROS) et le sur-échantillonnage synthétique de la classe minoritaire (SMOTE). L'état de l'art présente ces deux méthodes de sur-échantillonnage dans le cas de deux classes déséquilibrées (une classe majoritaire et une classe minoritaire). Pour trois classes déséquilibrées, le fonctionnement est le même, sauf que nous sommes en présence de deux classes minoritaires et d'une seule classe majoritaire. Les méthodes de sur-échantillonnage sont appliquées à toutes les classes minoritaires.
Afin de classifier nos données, nous utilisons une forêt d'arbres décisionnels qui est un algorithme proposant les meilleurs résultats. La forêt d'arbres décisionnels offre une meilleure explicabilité que les réseaux de neurones artificiels, facilitant ainsi la compréhension du modèle prédictif.

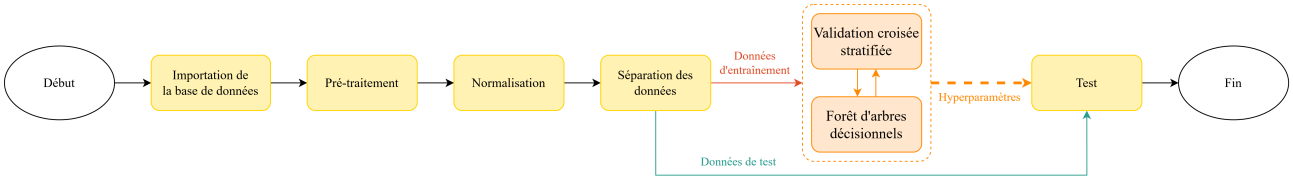

FIGURE 6 – Pipeline 1 (pas de sur-échantillonnage). Le cadre en pointillés orange correspond à l'optimisation des hyperparamètres.

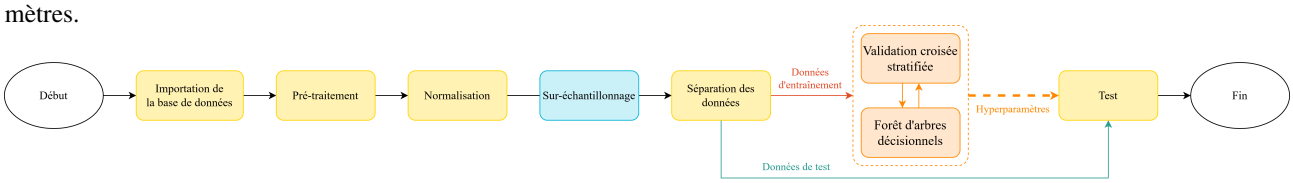

FIGURE 7 – Pipeline 2 (sur-échantillonnage avant la séparation des données). Le cadre en pointillés orange correspond à l'optimisation des hyperparamètres.

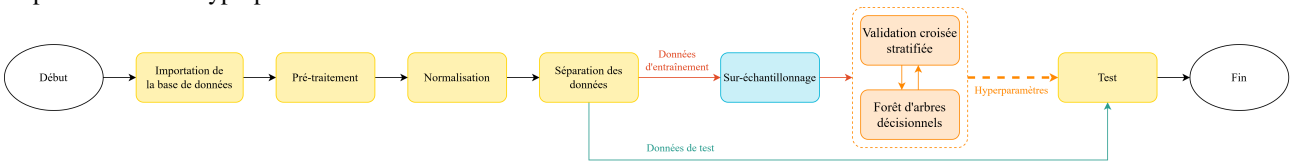

FIGURE 8 – Pipeline 3 (sur-échantillonnage après la séparation des données). Le cadre en pointillés orange correspond à l'optimisation des hyperparamètres.

Différents hyperparamètres sont testés pour déterminer la combinaison optimale :

- **Profondeur maximale d'un arbre** : 2, 3, 4, 5, 6, 7, 8 ou 9.
- **Nombre d'arbres** : 10, 25, 50 ou 100.
- **Critère** : Mesure d'impureté de Gini (BREIMAN [2]) ou mesure de l'entropie (QUINLAN [17]).

Le critère est une mesure utilisée pour évaluer la qualité de la division des nœuds dans un arbre de décision. Les formules 2 et 3 définissent respectivement l'impureté de Gini et l'entropie.

$$gini(Q_m) = \sum_k p_{mk}(1 - p_{mk}) \in [0, 0.5], \qquad (2)$$

où Q_m représente les données au nœud m et p_{mk} est la probabilité de la classe k à un nœud m.

$$entropie(Q_m) = -\sum_k p_{mk} log_2(p_{mk}) \in [0, 1], \qquad (3)$$

où Q_m représente les données au nœud m et p_{mk} est la probabilité de la classe k à un nœud m.

Pour évaluer les performances des modèles de prédiction, nous utilisons la validation croisée stratifiée à k blocs. Cette méthode consiste à tester le modèle sur différentes partitions, appelées blocs, de l'ensemble de données d'entraînement. Nous veillons à ce que chaque bloc contienne la même distribution de classes que les données d'entraînement (stratification).

Enfin, une fois que les hyperparamètres sont optimisés, les données de test sont présentées au modèle entraîné. Les précisions, les rappels et les F-scores des modèles sont calculés pour chaque classe selon les trois équations suivantes :

$$précision = \frac{TP}{TP + FP} \in [0, 1], \qquad (4)$$

où TP est le nombre de vrais positifs, FN est le nombre de faux négatifs et FP est le nombre de faux positifs.

$$rappel = \frac{TP}{TP + FN} \in [0, 1], \qquad (5)$$

où TP est le nombre de vrais positifs, FN est le nombre de faux négatifs et FP est le nombre de faux positifs.

$$F_1 = \frac{précision \times rappel}{précision + rappel} \in [0, 1], \qquad (6)$$

Nous calculons également l'étendue des scores selon l'équation 7. Une faible étendue des scores signifie qu'ils sont semblables entre les trois classes. Ces étendues seront comparées avec les scores maximaux et minimaux. Le modèle idéal possède une faible étendue et des valeurs maximales et minimales proches de 1.

$$étendue = |score_{max} - score_{min}| \qquad (7)$$

En résumé, le pipeline 1 est le modèle de référence. Le pipeline 2 est le modèle permettant de simuler le cas où les données issues de Motive sont équilibrées. Le pipeline 3 est le modèle que nous pourrions appliquer aux données réelles si celles-ci sont déséquilibrées.

4 Résultats

Dans cette partie, nous allons présenter l'ensemble des résultats obtenus avec les données de test. Nous parlerons du pipeline 1, puis du pipeline 2 et, enfin, du pipeline 3. Nous terminerons en comparant les trois pipelines.

4.1 Analyse des résultats du pipeline 1

Dans ce paragraphe, nous présentons les résultats du premier pipeline sans sur-échantillonnage.
La table 1 montre les scores moyens de précision, de rappel et de F-score pour chaque modèle testé sur différentes classes et chapitres. Les scores présentés sont la moyenne des scores pour les trois classes sur chacun des quatre chapitres (2A, 3A, 5A et 5B). Sans l'utilisation du sur-échantillonnage, le modèle a obtenu une précision moyenne de 0,75, un rappel moyen de 0,60 et un F-score moyen de 0,60.

Pipeline	Méthode	Précision	Rappel	F-score
1	-	0,75	0,60	0,60
2	ROS	**0,85**	**0,84**	**0,85**
	SMOTE	0,81	0,80	0,80
3	ROS	0,65	0,65	0,64
	SMOTE	0,64	0,65	0,64

TABLE 1 – Précisions, rappels et F-scores moyens des chapitres 2A, 3A, 5A et 5B. Les scores en gras correspondent aux plus grandes valeurs par colonne.

Les tables 2, 3 et 4 présentent les scores de précision, de rappel et de F-score pour chaque classe et pour les quatre chapitres. Pour le pipeline 1, le modèle a obtenu des précisions de 0,83 pour la classe Abandon, 0,69 pour la classe Succès et 0,73 pour la classe Échec. Cela signifie que le modèle a correctement prédit un nombre relativement élevé d'échantillons positifs par rapport à l'ensemble des échantillons prédits comme positifs.

Pipeline	Méthode	Abandon	Succès	Échec
1	-	0,83	0,69	0,73
2	ROS	**0,92**	**0,78**	**0,86**
	SMOTE	0,90	0,72	0,82
3	ROS	0,88	0,70	0,36
	SMOTE	0,88	0,70	0,33

TABLE 2 – Précisions moyennes des chapitres 2A, 3A, 5A et 5B. Les scores en gras correspondent aux plus grandes valeurs par colonne.

En ce qui concerne le rappel, le modèle a obtenu des scores de 0,84 pour la classe Abandon, 0,78 pour la classe Succès et 0,17 pour la classe Échec. Cela suggère que le modèle est généralement efficace pour détecter les abandons et les succès des étudiants, mais il a du mal à détecter les échecs. Cette difficulté est probablement due au déséquilibre des données.

Pipeline	Méthode	Abandon	Succès	Échec
1	-	**0,84**	0,78	0,17
2	ROS	0,80	**0,84**	**0,89**
	SMOTE	0,78	0,83	0,80
3	ROS	0,78	0,83	0,34
	SMOTE	0,78	0,79	0,39

TABLE 3 – Rappels moyens des chapitres 2A, 3A, 5A et 5B. Les scores en gras correspondent aux plus grandes valeurs par colonne.

Pipeline	Méthode	Abandon	Succès	Échec
1	-	0,83	0,73	0,25
2	ROS	**0,85**	**0,81**	**0,88**
	SMOTE	0,83	0,77	0,81
3	ROS	0,82	0,75	0,35
	SMOTE	0,82	0,73	0,36

TABLE 4 – F-scores moyens des chapitres 2A, 3A, 5A et 5B. Les scores en gras correspondent aux plus grandes valeurs par colonne.

En résumé, les résultats de ce pipeline montrent que le déséquilibre des données impacte les scores et nous allons analyser comment rééquilibrer les données pour améliorer nos résultats avec les pipelines 2 et 3.

4.2 Analyse des résultats du pipeline 2

Nous allons maintenant détailler les résultats obtenus pour le pipeline 2, qui implique l'utilisation du sur-échantillonnage avant la séparation des données. Deux méthodes de sur-échantillonnage sont utilisées : le sur-échantillonnage aléatoire (ROS) et le sur-échantillonnage synthétique de la classe minoritaire (SMOTE).
Avec la méthode de sur-échantillonnage aléatoire (ROS), le modèle obtient une précision moyenne de 0,85, un rappel moyen de 0,84 et un F-score moyen de 0,85. Les scores de précision pour les trois classes varient entre 0,78 et 0,92, les scores de rappels varient entre 0,80 et 0,89 et les scores de F-score varient entre 0,81 et 0,88.
Les résultats obtenus avec le sur-échantillonnage synthétique de la classe minoritaire (SMOTE) sont légèrement inférieurs aux résultats du sur-échantillonnage aléatoire (ROS). Le modèle obtient une précision moyenne de 0,81, un rappel moyen de 0,80 et un F-score moyen de 0,80. Les scores de précision pour les trois classes varient entre 0,72 et 0,90, les scores de rappel varient entre 0,78 et 0,83 et les

scores de F-score varient entre 0,77 et 0,83.
En comparant les deux méthodes de sur-échantillonnage, le sur-échantillonnage aléatoire (ROS) obtient de meilleurs scores moyens que le sur-échantillonnage synthétique de la classe minoritaire (SMOTE) lorsque le sur-échantillonnage s'effectue avant la séparation des données. La méthode SMOTE créée des échantillons synthétiques à partir des échantillons originaux. Cette méthode de sur-échantillonnage semble introduire du bruit dans les données et, par conséquent, diminuer les performances des modèles de prédiction.

4.3 Analyse des résultats du pipeline 3

Enfin, nous présentons les résultats du pipeline 3 qui implique l'utilisation du sur-échantillonnage après la séparation des données.
Avec la méthode de sur-échantillonnage aléatoire (ROS), le modèle obtient une précision moyenne de 0,65, un rappel moyen de 0,65 et un F-score moyen de 0,64. Les scores de précision pour les trois classes varient entre 0,36 et 0,88, le scores de rappel varient entre 0,34 et 0,83 et les scores de F-scores varient entre 0,35 et 0,82.
Les résultats obtenus avec le sur-échantillonnage synthétique de la classe minoritaire (SMOTE) sont encore une fois légèrement inférieurs aux résultats du sur-échantillonnage aléatoire (ROS). Le modèle obtient une précision moyenne de 0,64, un rappel moyen de 0,65 et un F-score moyen de 0,64. Les scores de précision pour les trois classes varient entre 0,33 et 0,88, les scores de rappel varient entre 0,39 et 0,79 et les scores de F-score varient entre 0,36 et 0,82.
En comparant les deux méthodes de sur-échantillonnage, le sur-échantillonnage aléatoire (ROS) obtient des scores moyens comparables à ceux du sur-échantillonnage synthétique de la classe minoritaire (SMOTE) lorsque le sur-échantillonnage s'effectue après la séparation des données. Cependant, les scores obtenus avec le pipeline 3 sont inférieurs aux scores obtenus avec le pipeline 2.

4.4 Comparaison des pipelines

En comparant les trois pipelines, nous constatons que les meilleurs résultats sont obtenus avec le pipeline 2, lorsqu'un sur-échantillonnage aléatoire est effectué avant la séparation des données. Les prédictions sont donc meilleures lorsque les données sont équilibrées artificiellement puis séparées pour entraîner le modèle. Cependant, les données de test de ce pipeline ne sont pas représentatives d'une population réelle. Le pipeline 3 possède des données de test représentatives d'une population réelle. Il obtient des rappels moyens et F-scores moyens supérieurs, mais les précisions moyennes sont inférieures à celle du pipeline 1 (pour les deux méthodes de sur-échantillonnage). La présence du sur-échantillonnage semble donc diminuer les performances du modèle. La table 5 présente les précisions moyennes obtenues par le modèle à l'issue de l'entraînement et à l'issue du test. Le pipeline 1 obtient une précision moyenne d'entraînement de 0,84. Le pipeline 3 obtient une précision moyenne d'entraînement de 0,93 (ROS) et 0,91 (SMOTE). Le sur-échantillonnage des données d'entraînement entraîne donc un phénomène de sur-apprentissage des données. Le modèle du pipeline 3 généralise moins bien et possède donc une précision inférieure à celle du pipeline 1. Ce sur-apprentissage se constate sous une autre forme : le rappel moyen de la classe Abandon du pipeline 1 est supérieur aux rappels moyens de la même classe du pipeline 3 (pour les deux méthodes de sur-échantillonnage). Nous remarquons également que la précision moyenne de la classe Échec est plus élevée avec le pipeline 1 qu'avec le pipeline 3 (pour les deux méthodes de sur-échantillonnage).

Pipeline	Méthode	Entraînement	Test
1	-	0,84	0,75
2	ROS	0,91	0,85
	SMOTE	0,90	0,81
3	ROS	0,93	0,65
	SMOTE	0,91	0,64

TABLE 5 – Précisions moyennes des chapitres 2A, 3A, 5A et 5B lors de l'entraînement et lors du test du modèle.

L'étendue des scores est également à prendre en compte. En observant la table 6, nous constatons que l'étendue des scores est améliorée dans le cas du pipeline 2 par rapport aux deux autres pipelines. L'étendue des scores obtenus dans ce cadre est toujours inférieure à 0,2, ce qui montre une faible dispersion des résultats en fonction des classes et des chapitres et prouve que le sur-échantillonnage avant la séparation des données est le plus efficace pour prédire, de manière fiable, les résultats de étudiants.

Pipeline	Méthode	Précision	Rappel	F-score
1	-	0,15	0,67	0,59
2	ROS	**0,14**	0,09	**0,07**
	SMOTE	0,18	**0,05**	**0,07**
3	ROS	0,51	0,49	0,47
	SMOTE	0,54	0,40	0,46

TABLE 6 – Étendues des scores. Les scores en gras correspondent aux plus petites valeurs par colonne.

Finalement, la figure 7 donne une vision globale des résultats. Le sur-échantillonnage avant la séparation des données (pipeline 2) permet d'améliorer les scores du modèle de prédiction. Il obtient aussi les plus petites étendues des scores. La stratégie de sur-échantillonnage du pipeline 3 n'est pas assez performante pour obtenir des résultats satisfaisants. Notons également que cette étude est généralisable à l'ensemble des chapitres de Motive. En calculant les précisions, rappels et F-scores moyens de tous les chapitres de Motive, nous obtenons des valeurs comparables aux résultats de la table 2. Les étendues des scores sont légèrement supérieures à celles de la figure 9.

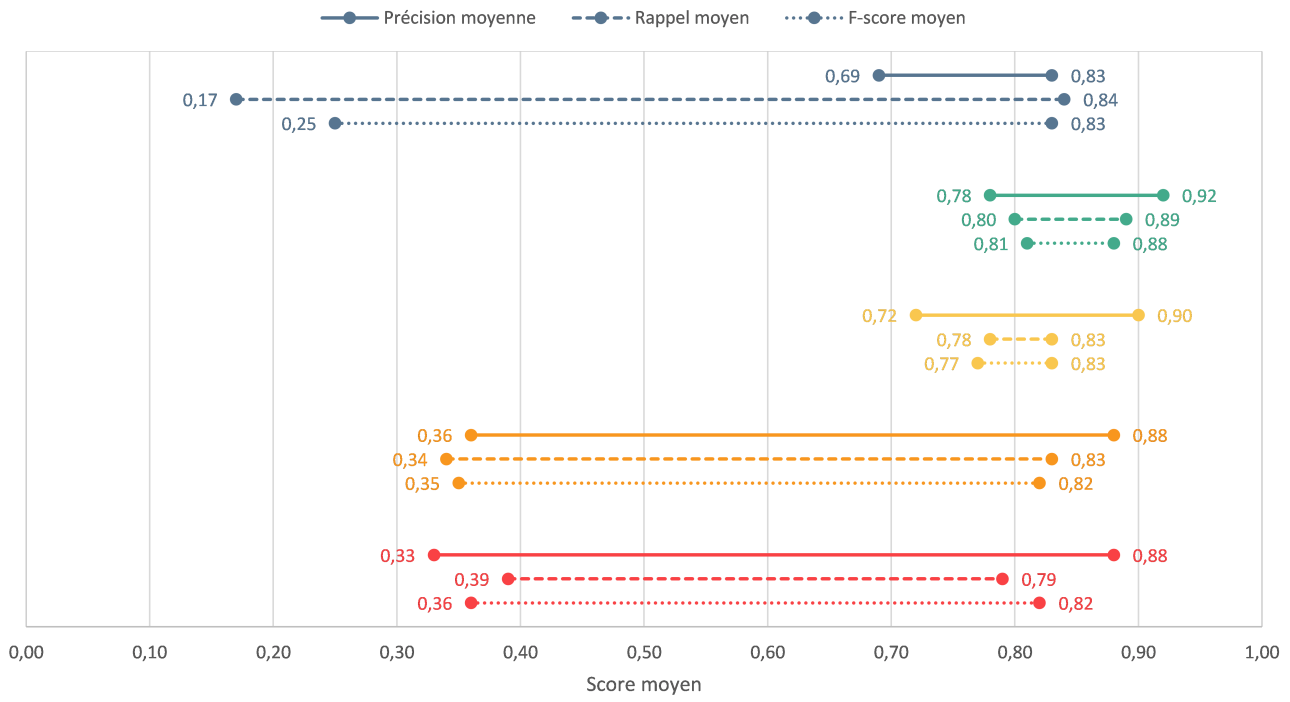

FIGURE 9 – Valeurs minimales et maximales des scores moyens des chapitres 2A, 3A, 5A et 5B. De haut en bas : en bleu, le pipeline 1 ; en vert, le pipeline 2 (ROS) ; en jaune, le pipeline 2 (SMOTE) ; en orange, le pipeline 3 (ROS) et en rouge, le pipeline 3 (SMOTE).

Notre objectif était de prédire des profils étudiants à partir des données issues de la plateforme d'apprentissage en ligne Motive. Ces données présentaient un déséquilibre entre les trois classes. Le sur-échantillonnage a permis d'améliorer les performances des modèles de prédiction de manière significative. Notre contribution réside dans l'utilisation de données (réelles et déséquilibrées) issues uniquement de la plateforme Motive. Aucune information sur les étudiants telles que la provenance géographique, l'historique scolaire et universitaire ou encore le niveau de stress n'ont été utilisé.

5 Conclusion et perspectives

L'objectif de ces travaux était de détecter des profils étudiants à partir des traces numériques des étudiants sur la plateforme Motive. Le principal problème de ces données était le déséquilibre des classes. La classe Abandon est beaucoup plus représentée que les classes Succès et Échec. Pour lutter contre ce déséquilibre, deux méthodes de sur-échantillonnage ont été testées : le sur-échantillonnage aléatoire (ROS) et le sur-échantillonnage synthétique de la classe minoritaire (SMOTE). Les méthodes ont été testées avant et après la séparation des données. Les résultats montrent que le sur-échantillonnage aléatoire avant la séparation des données est la meilleure méthode pour prédire au mieux les profils étudiants. Elle permet d'augmenter la précision, le rappel et F-score. Par ailleurs, le rappel de la classe Échec a considérablement augmenté grâce au sur-échantillonnage.
En utilisant les données d'entraînement des étudiants et en augmentant artificiellement le nombre d'échantillons, il est donc possible de prédire l'abandon, le succès ou l'échec de l'étudiant à un chapitre de Motive. Outre la notion de succès, ces prédictions peuvent permettent aux enseignants et aux professeurs de l'enseignement supérieur de détecter les étudiants en situation de décrochage scolaire, ou ceux en difficultés.
La suite de ces travaux portera sur l'analyse du Moodle complet de l'Université de Poitiers. Le principal défi réside dans l'étiquetage des données et la proposition d'indicateur pertinent dans le suivi de l'étudiant. En parallèle, la publication d'une base de données, similaire à celle de Motive, pourrait être bénéfique afin de permettre à la communauté scientifique de l'explorer, d'effectuer des expérimentations et de comparer les résultats obtenus.

Références

[1] S. K. BANIHASHEM et al. "A systematic review of the role of learning analytics in enhancing feedback practices in higher education". In : *Educational Research Review* 37 (2022), p. 100489.

[2] L. BREIMAN, éd. *Classification and regression trees*. 1998.

[3] L. BREIMAN. "Random forests". In : *Machine Learning* 45.1 (2001), p. 5-32.

[4] N. V. CHAWLA et al. “SMOTE : Synthetic Minority Over-sampling Technique”. In : *Journal of Artificial Intelligence Research* 16 (2002), p. 321-357.

[5] D. FOUNG et J. CHEN. “A Learning Analytics Approach to the Evaluation of an Online Learning Package in a Hong Kong University”. In : *Electronic Journal of e-Learning* 17.1 (2019).

[6] B. K. FRANCIS et S. S. BABU. “Predicting Academic Performance of Students Using a Hybrid Data Mining Approach”. In : *Journal of Medical Systems* 43.6 (2019), p. 162.

[7] H. HAN, W.-Y. WANG et B.-H. MAO. “Borderline-SMOTE : A New Over-Sampling Method in Imbalanced Data Sets Learning”. In : *Advances in Intelligent Computing*. T. 3644. 2005, p. 878-887.

[8] H. HASSAN, N. B. AHMAD et S. ANUAR. “Improved students’ performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining”. In : *Journal of Physics : Conference Series* 1529.5 (2020), p. 052041.

[9] O. IATRELLIS et al. “A two-phase machine learning approach for predicting student outcomes”. In : *Education and Information Technologies* 26.1 (2021), p. 69-88.

[10] J. KUZILEK, M. HLOSTA et Z. ZDRAHAL. “Open University Learning Analytics dataset”. In : *Scientific Data* 4.1 (2017), p. 170171.

[11] J. KUZILEK et al. “Analysing Student VLE Behaviour Intensity and Performance”. In : *Transforming Learning with Meaningful Technologies*. 2019, p. 587-590.

[12] J. MACQUEEN. “Some Methods For Classification And Analysis Of Multivariate Observations”. In : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (1967).

[13] N. C. MING et V. L. MING. “Predicting Student Outcomes from Unstructured Data”. In : (2012).

[14] P. M. MORENO-MARCOS et al. “Analysis of the Factors Influencing Learners’ Performance Prediction With Learning Analytics”. In : *IEEE Access* 8 (2020), p. 5264-5282.

[15] A. NAMOUN et A. ALSHANQITI. “Predicting Student Performance Using Data Mining and Learning Analytics Techniques : A Systematic Literature Review”. In : *Applied Sciences* 11.1 (2020), p. 237.

[16] H. M. NGUYEN, E. W. COOPER et K. KAMEI. “Borderline over-sampling for imbalanced data classification”. In : *International Journal of Knowledge Engineering and Soft Data Paradigms* 3.1 (2011), p. 4.

[17] J. R. QUINLAN. *C4.5 : programs for machine learning*. The Morgan Kaufmann series in machine learning. Morgan Kaufmann Publishers, 1993. 302 p.

[18] T. WONGVORACHAN, S. HE et O. BULUT. “A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining”. In : *Information* 14.1 (2023), p. 54.

Session 5 : Approches numériques-symboliques

Approche incrémentale pour la détection des textes de légendes dans des cartes numériques

A. Marzinkowski[1], S. Benferhat[1], A. Paparrizou[2], C. Piette[1]

[1] Centre de Recherche en Informatique de Lens, CRIL
[2] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, LIRMM

{marzinkowski,benferhat,piette}@cril.fr paparrizou@lirmm.fr

Résumé

Cet article concerne la détection automatique des textes de légende dans les cartes. Après avoir extrait les textes des images, avec des outils OCR, nous utiliserons un processus itératif de regroupement ("clustering") des textes extraits. Cinq critères principaux, avec différents niveaux d'importance, sont utilisés : l'alignement des textes, la distance entre les zones de texte, la couleur de fond des textes, la couleur des textes et la taille des caractères. Pour chacun des critères, des mesures de similarité appropriées sont définies. Nous proposons une méthode qui combinerait de manière hiérarchique les regroupements obtenus à partir de chaque critère. L'étude expérimentale révèle deux résultats importants. Premièrement, l'utilisation de plusieurs critères donne des résultats supérieurs à ceux d'une simple distance (e.g., euclidienne) entre les zones de texte. Deuxièmement, l'étude expérimentale confirme l'efficacité globale de la relation de priorité que nous avons intuitivement définie entre les critères pour détecter à partir des textes de la légende.

Mots-clés

Cartes numériques, OCR, clustering

Abstract

This paper deals with the automatic detection of legend texts inside maps. After extracting the texts from the images, using OCR tools, we make use of an iterative clustering process of the extracted texts. We. consider five main criteria, with different levels of importance : text alignment, distance between text boxes, background color of the text, font color and font size. For each criterion, we define appropriate similarity measures. We propose a method that combines, in an incremental way, the partitions obtained by each criterion. The experimental study reveals two important results. First, the combination of several criteria gives better results than just considering a simple distance (e.g., Euclidean distance) between text boxes. Secondly, the overall effectiveness of the priority relation, which we have intuitively defined between the criteria for detect with caption texts, is confirmed.

Keywords

Digital maps, OCR, clustering

1 Introduction

La détection automatique de textes à l'intérieur d'images est étudiée par la communauté de la vision par ordinateur depuis plusieurs décennies. La détection de texte et le regroupement de texte est principalement basée sur des caractéristiques visuelles extraites de l'image (c'est-à-dire la couleur, la forme, etc.) [5]. Ce travail concerne la détection de texte dans des images appartenant à la légende d'une carte numérique arbitraire, et n'est donc pas lié à un domaine d'information spécifique. La nouveauté de ce travail est que non seulement des caractéristiques visuelles sont utilisées, mais également des notions sémantiques qui accompagnent et facilitent la détection des zones de légende. Par exemple, une information sémantique peut être : la position possible de l'objet sur la carte, la relation ou la distance entre les objets voisins, etc.

Les images que nous traitons sont des images qui représentent des cartes numériques. Ce travail constitue une étape primordiale dans la détection d'objets d'intérêt à l'intérieur des cartes, qui sont généralement définis et affichés dans la légende. Pour ce faire, nous commençons par utiliser un algorithme de reconnaissance optique de caractères (OCR) afin d'obtenir les zones de texte qui apparaissent à l'intérieur d'une carte donnée. Nous utilisons les régions de texte fournies par l'algorithme OCR comme entrées dans notre algorithme de regroupement (*clustering*) k-means [16]. Nous introduisons également des critères qui caractérisent de manière symbolique ou visuelle la notion de légende ; par exemple, des phrases de texte alignées verticalement, pouvant former une légende. Nous définissons cinq de ces critères ainsi que les mesures de distance associées à chaque critère pour l'algorithme des k-means. Nous déployons la méthode des k-moyennes de manière incrémentale, où les *clusters* sont construits en fonction d'un critère pris en compte à chaque itération. L'ordre dans lequel les critères sont considérés joue donc un rôle crucial dans l'efficacité de notre algorithme. À chaque itération, notre algorithme doit décider s'il faut ou non diviser (ou éclater) chaque "cluster" obtenu à l'itération précédente. Cette déci-

sion est basée sur l'entropie que nous utilisons pour refléter l'homogénéité d'une partition.

Nous évaluons les résultats de la détection de la légende sur plusieurs cartes de tailles et de styles différents à l'aide de deux mesures d'évaluation. Les résultats montrent que notre algorithme détecte bien la région de la légende sur plusieurs cartes. L'analyse de notre étude exhaustive des critères indique que lorsque plusieurs critères sont pris en compte, nous obtenons une meilleure détection de la région de la légende. D'autres observations seront présentées dans la section consacrée à l'étude expérimentale.

Le reste de l'article est organisé comme suit. La section suivante présente brièvement certains travaux existants. La section 3 décrit le problème considéré dans cet article, définit les cinq critères utilisés pour le regroupement et présente notre algorithme de détection du texte de la légende. La section 4 contient les résultats des études expérimentales réalisées et enfin, la section 5 conclut cet article.

2 Travaux antérieurs

La détection ou la reconnaissance de texte concerne généralement les documents texte et de nombreux algorithmes OCR ont été proposés à cet effet avec une excellente précision [11]. Ces dernières années, la détection de texte dans les images de scènes naturelles a suscité beaucoup d'intérêt en raison d'une variété d'applications : compréhension de scènes, récupération d'images basées sur le contenu, aide à la navigation (pour les voitures autonomes ou les personnes malvoyantes) [1]. Dans [17], les auteurs ont proposé un cadre de détection de chaînes de texte, dans lequel le regroupement des caractères candidats est basé sur des caractéristiques structurelles, telles que les différences de taille des caractères, les distances entre les caractères voisins et l'alignement des caractères. Leur méthode de regroupement de lignes de texte effectue une transformation de Hough pour ajuster la ligne de texte aux centroïdes des candidats de texte au lieu de les regrouper. Ce travail est similaire au nôtre dans le sens où les notions sémantiques sont prises en compte, mais l'objectif et la granularité sont différents (regroupement de caractères et de lignes, et non regroupement de zones de texte). De même dans [6], le système d'assistance proposé reconnaît le texte dans une image en fonction de caractéristiques structurelles telles que la taille, l'orientation et la distance entre les régions d'intérêt successives.

Dans la communauté de la recherche d'informations, il y a un effort considérable pour le regroupement de texte avec des k-means, mais à des fins différentes des nôtres (c'est-à-dire, pour noter et classer la pertinence d'un document compte tenu d'une requête utilisateur [3], pour regrouper des documents de contenu similaire [8] ou pour résumer du texte [7]). Il existe des travaux sur la détection de texte dans les cartes *raster*, mais ils sont davantage liés aux approches OCR, où le défi est dû aux différentes orientations du texte et au chevauchement des étiquettes de texte [2, 9]. À notre connaissance, il n'existe aucun travail traitant de la détection de texte appartenant aux légendes des cartes.

3 Présentation du problème et critères de regroupement

3.1 Présentation du problème

Le problème que nous cherchons à résoudre est celui d'identifier la région d'une carte où se situent les textes de légende. Les entrées de notre algorithme sont des images qui représentent des cartes avec des légendes. Les cartes sur lesquelles nous travaillons sont constituées de figures mais aussi de zones de texte. Dans cet article, nous proposons une approche pour discriminer le texte appartenant à une légende potentielle des autres régions de texte de l'image.

Chaque image sera représentée par une matrice $n \times m$ éléments, que nous noterons par la suite par $\mathscr{I}$. Chaque élément de la matrice, représente une couleur d'un pixel (représentée ici dans le format RGB ; c'est-à-dire un triplet d'entiers compris entre 0 et 255). Nous utiliserons aussi x_i pour désigner un numéro d'un ligne et y_j pour désigner le numéro d'une colonne de la matrice $\mathscr{I}$.
Dans cet article, nous nous concentrons sur les images contenant des légendes. En particulier, nous nous intéressons à la zone de la carte qui contient le texte de la légende et qui sera également représentée par une matrice de couleurs de pixels notée $\mathscr{L}$. La sortie de notre algorithme est une zone de la matrice $\mathscr{I}$ qui est censée représenter la zone de texte de la légende $\mathscr{L}$.
Nous présentons maintenant deux notions qui seront utilisées plus tard par notre algorithme. La première notion, que nous appelons éclatement, consiste à construire une partition d'un ensemble à partir d'une partition plus grande. La seconde est un rappel de la notion de l'entropie qui servira à mesurer l'homogénéité d'une partition.

Definition 1 (Eclatement) *Soit A un ensemble d'éléments. Soit B un sous-ensemble de A et $\mathscr{P}_A$ une partition de A. Nous appelons l'éclatement de B par $\mathscr{P}_A$, noté $B \triangleright \mathscr{P}_A$, la partition de B obtenue en intersectant chaque élément de $\mathscr{P}_A$ avec B. Plus formellement :*

$$B \triangleright \mathscr{P}_A = \{B \cap C_i : C_i \in \mathscr{P}_A\} \tag{1}$$

Definition 2 (Mesure d'homogénéité) *Soit A un ensemble d'éléments et $\mathscr{P}_A$ une partition de A. Nous définissons l'homogénéité (ou l'entropie) de $\mathscr{P}_A$, notée $\mathscr{E}(\mathscr{P}_A)$, par :*

$$\mathscr{E}(\mathscr{P}_A) = -\sum_{B_i \in \mathscr{P}_A} \frac{\|B_i\|}{\|A\|} * \log_2 \frac{\|B_i\|}{\|A\|}, \tag{2}$$

où $\|x\|$ représente la cardinalité de x.

3.2 Extraction des textes depuis les cartes

La première étape de notre algorithme consiste à extraire des textes depuis les cartes. Dans cette étape, nous utiliserons simplement les outils OCR existants. Nous utilisons notamment l'outil OCR DocTr [10] qui permet d'obtenir à partir d'une image une liste de zones de texte avec

différents niveaux de granularité (des blocs de texte, une ligne de texte, un mot, etc.).

A partir d'une carte $\mathscr{I}$, l'outil OCR retourne un ensemble de texte, noté $\mathscr{T}_{\mathscr{I}}$. Les éléments de $\mathscr{T}_{\mathscr{I}}$ sont appelés des boîtes de textes et sont dénotés par les lettres calligraphiques minuscules $\mathscr{a}, \mathscr{b}, \mathscr{c}, etc.$

La Figure 1 illustre un exemple de zone de texte $\mathscr{a}$ qui est représentée par un rectangle sur la carte identifié par deux points aux extrémités d'une diagonale $(x^{\mathscr{a}}, y^{\mathscr{a}})$ et $(x_{\mathscr{a}}, y_{\mathscr{a}})$.

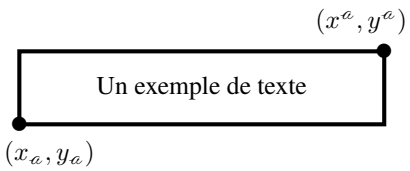

FIGURE 1 – Un exemple de boîte de texte représentée par un rectangle

Remarques :

— Dans le reste de cet article, nous supposons que tous les rectangles associés aux boîtes de texte sont disjoints.

— Nous excluons aussi les rectangles qui ne sont pas verticaux ou horizontaux (par rapport aux axes de l'image) du fait que les légendes ne sont pas inclinés.

3.3 Définitions des critères de regroupement

Cette sous-section présente les cinq principaux critères qui seront utilisés par notre méthode de détection des zones de texte de légende. Pour chacun des cinq critères, nous définissons les mesures de similarité associées pour comparer deux zones de texte. Soit $\mathscr{a}$ et $\mathscr{b}$ deux zones de texte.
Par abus de langage, nous utiliserons le terme distance pour exprimer la similarité entre deux zones de texte, sans exiger de propriété préalable sur la mesure de distance utilisée.

Critère 1 : l'alignement des textes

Le texte d'une légende est souvent aligné. L'alignement du texte peut prendre différentes formes : à gauche, au centre, à droite, etc. Par souci de simplicité, nous nous limitons uniquement à la situation des légendes avec un texte aligné verticalement à gauche. L'extension à d'autres formes d'alignement se fait facilement par symétrie ou par une légère adaptation de la mesure de distance définie ci-dessous.
La distance associée à l'alignement gauche, notée d_{ag}, est définie par :

$$d_{ag}(\mathscr{a}, \mathscr{b}) = \mid y_{\mathscr{a}} - y_{\mathscr{b}} \mid . \quad (3)$$

Critère 2 : la distance entre les textes

Un deuxième critère naturel est celui de la mesure de similarité entre les zones de texte que nous noterons d_e. Les textes des légendes sont en général proches les uns des autres.

Une première idée pour définir la distance entre deux boîtes de texte est de considérer la plus petite distance entre deux points quelconques des périmètres de rectangles associés aux deux boîtes de texte. Cette solution n'est pas satisfaisante dans notre cadre car nous utilisons des rectangles particuliers (disjoints, horizontaux ou verticaux, texte alignés à gauche, etc.).

La définition de la distance entre deux boîtes dépend clairement des résultats des OCR qui peuvent contenir des phrases entières ou simplement des mots. Nous proposons d'analyser chacune de ces deux situations.
Commençons par le cas où les boîtes contiennt des phrases entières. La figure 2 donne la situation "idéale" dans laquelle la distance entre deux boîtes $\mathscr{a}$ et $\mathscr{b}$ est égale à 0. Il s'agit de deux boîtes parfaitement alignées à gauche (nous rappelons pour des raisons de simplicité seul l'alignement à gauche est considérée) et dont la distance est égale à l'unité (un seul pixel dans notre cas).

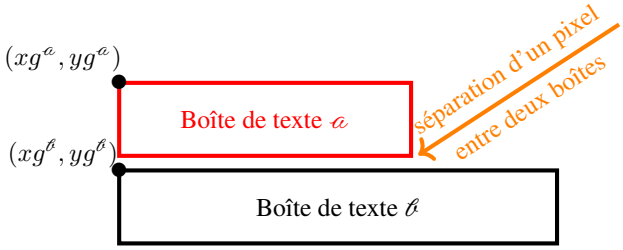

FIGURE 2 – Situations où deux boîtes sont considérées le plus proches possible

Plus formellement, soit $\mathscr{a}$ ey $\mathscr{b}$ deux boîtes de texte disjointes alors :

$$d_e(\mathscr{a}, \mathscr{b}) = 0$$
$$\text{ssi}$$
$$(y_{\mathscr{a}} = y_{\mathscr{b}})$$
$$\text{et}$$
$$[(xg^{\mathscr{a}} = x_{\mathscr{b}} + 1) \text{ ou } (xg^{\mathscr{b}} = x_{\mathscr{a}} + 1)].$$

Notons que la distance ne s'applique qu'avec deux boîtes de texte disjointes (c'est le cadre de notre article).
Sur la base de cette définition de situations où deux boîtes de texte sont considérées comme idéalement proches, il suffit alors de définir la distance comme la tranlation nécessaire d'une des deux boîtes pour atteindre cette situation idéale.
Plus formellement, à partir des notations données dans la figure 3, nous avons :

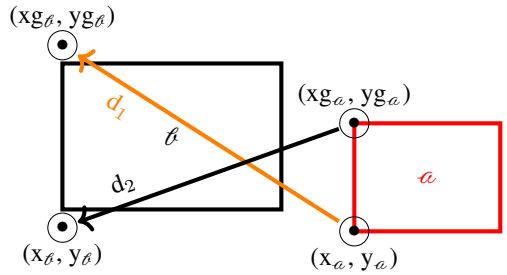

FIGURE 3 – Illustration du calcul de la distance entre deux boîtes de texte

$$\begin{aligned} d_e(\mathscr{a},\mathscr{b}) &= min(d_1,d_2), \text{avec} \\ d_1 &= \sqrt{(x_{\mathscr{a}} - xg_{\mathscr{b}})^2 + (y_{\mathscr{a}} - yg_{\mathscr{b}})^2}, et \\ d_2 &= \sqrt{(x_{\mathscr{b}} - xg_{\mathscr{a}})^2 + (y_{\mathscr{b}} - yg_{\mathscr{a}})^2}. \end{aligned} \quad (4)$$

Bien sûr, d'autres mesures peuvent être utilisées (comme la distance de Manhattan) à la place de la distance euclidienne.

Maintenant dans l'hypothèse où les boîtes de texte contiennent uniquement des mots, dans ce cas la distance (euclidienne) entre les centres des mots est suffisante. Notons $(xc_{\mathscr{a}}, yc_{\mathscr{a}})$ et $(xc_{\mathscr{b}}, yc_{\mathscr{b}})$ les coordonnées des points qui représentent les centres des boîtes $\mathscr{a}$ et $\mathscr{b}$ respectivement. La distance entre $\mathscr{a}$ et $\mathscr{b}$ est simplement définie par :

$$d_e(\mathscr{a},\mathscr{b}) = \sqrt{(xc_{\mathscr{a}} - xc_{\mathscr{b}})^2 + (yc_{\mathscr{a}} - yc_{\mathscr{b}})^2} \quad (5)$$

Critère 3 : la couleur des fonds des boîtes de texte

Le troisième critère est la couleur de fond de la zone de texte. En effet, dans les légendes le fond utilisé est souvent de couleur homogène. L'utilisation de ce critère soulève deux questions, heureusement bien abordées dans la littérature sur le traitement d'images (e.g., [12]).

La première question est de savoir comment déterminer l'arrière-plan d'une boîte de texte $\mathscr{a}$. Dans notre contexte, on peut simplement utiliser la fréquence de couleur des pixels car on peut raisonnablement supposer que les textes dans les cases à comparer sont de couleur homogène et que les textes occupent la majorité de la case. Une autre façon de procéder, qui est celle utilisée dans l'article, est de calculer la partie la plus connectée (connexes) de la boite (deux pixels sont connectés s'ils sont de couleur homogène, c'est-à-dire que les deux pixels sont de couleur suffisamment similaire) en partant de l'un des bords de la boîte de texte.

La deuxième question est de savoir comment déterminer la couleur représentative du fond du texte. Là encore, nous avons plusieurs possibilités (e.g. [15]). Une possibilité à partir des valeurs de RGB de chaque pixel de la zone de texte, est de i) calculer d'abord la somme des carrées des couleurs (valeur par valeur), iii) puis diviser le résultat par le nombre de pixels et iii) enfin appliquer la racine carrée au résultat (pour rester dans l'ensemble $\{0, ..., 255\}$). Une autre possibilité est tout simplement de prendre la moyenne des couleurs (valeur par valeur) des différents pixels de du fond de l'image. C'est cette méthode qui est utilisée dans cet article.

Notons maintenant $(RF_{\mathscr{a}}, GF_{\mathscr{a}}, BF_{\mathscr{a}})$ et $(RF_{\mathscr{b}}, GF_{\mathscr{b}}, BF_{\mathscr{b}})$ les couleurs (en moyenne) du fond de texte des boîtes de texte $\mathscr{a}$ et $\mathscr{b}$ respectivement. Reste maintenant à calculer la proximité entre ces deux couleurs où plusieurs définitions sont possibles. Dans cet article, nous utilisons la distance euclidienne, notée d_{bg}, et définie par :
$d_{bg}(\mathscr{a},\mathscr{b}) =$

$$\sqrt{(RF_{\mathscr{a}} - RF_{\mathscr{b}})^2 + (GF_{\mathscr{a}} - GF_{\mathscr{b}})^2 + (BF_{\mathscr{a}} - BF_{\mathscr{b}})^2}. \quad (6)$$

Critère 4 : la hauteur des boîtes de texte

Les trois critères décrits ci-dessus (alignement, distance et fond de la zone de texte) sont fondamentaux pour déterminer la zone de texte d'une légende. Un autre critère concerne la taille de la police des caractères utilisée qui est ici supposée égale à la hauteur de la zone de texte.
La quatrième distance, notée d_t, donnée par la hauteur des boîtes de texte est simplement définie par :

$$d_t(\mathscr{a},\mathscr{b}) = \mid (\mid x_{\mathscr{a}} - x^{\mathscr{a}} \mid) - (\mid x_{\mathscr{b}} - x^{\mathscr{b}} \mid) \mid \quad (7)$$

Critère 5 : la couleur du texte

Ce dernier critère est le dual du critère 3 (la couleur du fond d'une boîte de texte), puis que l'on considère qu'une boîte de texte peut-être divisée en deux partie : la partie qui contient les caractères de la boîte de texte et le reste qui représente le fond de la boîte de texte. Notons $(RT_{\mathscr{a}}, GT_{\mathscr{a}}, BT_{\mathscr{a}})$ et $(RT_{\mathscr{b}}, GT_{\mathscr{b}}, BT_{\mathscr{b}})$ les couleurs (en moyenne) des textes des boîtes de texte $\mathscr{a}$ et $\mathscr{b}$ respectivement. Alors la distance par rapport à la couleur des textes, notée d_{ct}, est définie par :
$d_{ct}(\mathscr{a},\mathscr{b}) =$

$$\sqrt{(RF_{\mathscr{a}} - RT_{\mathscr{b}})^2 + (GT_{\mathscr{a}} - GT_{\mathscr{b}})^2 + (BT_{\mathscr{a}} - BT_{\mathscr{b}})^2}. \quad (8)$$

3.4 Algorithme par raffinements successifs des résultats de regroupement

Nous avons choisi une approche à base de *clustering* [18, 4]. Comme nos boîtes de texte extraites depuis les images ne sont pas étiquetées, nous utiliserons des méthodes basées sur l'apprentissage non-supervisée. Plus précisément, dans cet article nous avons opté pour l'algorithme k-means (e.g., [16]) largement utilisé dans des problèmes de classification non-supervisé.
Nous cherchons à regrouper des régions de texte de telle manière que ces regroupements représentent le texte d'une

légende. Les données d'entrée de notre algorithme sont avant tout une carte dont nous supposons qu'elle contient une légende. Cette carte sera notée $\mathscr{I}$. A ces données, nous ajoutons deux paramètres. Le premier est un tableau, noté $\vec{c}$, de critères (de taille 1 à 5) qui indique l'ordre dans lequel les critères doivent être utilisés.
Le deuxième paramètre est un fonction seuil, noté $\sigma(\mathscr{A})$ qui indique si une partition d'un ensemble $\mathscr{A}$ est suffisamment homogène ou non. Comme pour toute fonction de collecte, la question difficile est de savoir comment fixer ce seuil. Dans notre étude expérimentale, il est fixé à 80 % de la valeur maximale.

A ces deux paramètres s'ajoutent deux autres fonctions. La première $OCR(\mathscr{I})$ qui retourne $\mathscr{T}_\mathscr{I}$ l'ensemble des boîtes de texte qui se trouvent dans l'image $\mathscr{I}$). La deuxième fonction est la méthode de regroupement utilisée. Comme nous l'avons indiqué plus haut, nous utilisons simplement la méthode de k-means.

Algorithm 1 Algorithme de détection de légendes

Entrées :
- $\mathscr{I}$: une carte avec une légende
- $\vec{c}$: un vecteur de $n \in \{1, ..., 5\}$ critères
- σ : une fonction qui prend une partition et qui retourne un réel
- k-means : la méthode de regroupement utilisée

Sorties :
- $\mathscr{P}$ Un ensemble de *clusters* .

1: *// D'abord l'outil OCR est appliqué sur l'image*
2: $\mathscr{T}_\mathscr{I} \leftarrow OCR(\mathscr{I})$
3: *// On applique l'algorithme k-means sur* $\mathscr{T}_\mathscr{I}$
4: *// et le premier critère* $\vec{c}[1]$
5: $\mathscr{P} \leftarrow k - means(\mathscr{T}_\mathscr{I}, \vec{c}[1])$
6: *// On raffine l'ensemble* $\mathscr{P}$ *itérativement avec les autres critères*
7: **for all** $i \in \{2, ..., n\}$ **do**
8: *// On applique l'algorithme K-means sur* $\mathscr{T}_\mathscr{I}$
9: *// et le ième critère* $\vec{c}[i]$
10: $\mathscr{C} \leftarrow k - means(\mathscr{T}_\mathscr{I}, \vec{c}[i])$
11: *// On raffine chacun des éléments de*
12: *// la partition courante* $\mathscr{P}$
13: *// On stocke les résultats dans* $\mathscr{X}$
14: $\mathscr{X} \leftarrow \emptyset$
15: **for all** $B \in \mathscr{P}$ **do**
16: *// On éclate B* $\mathscr{C}$ *en utilisant Definition 1*
17: $R_B = B \triangleright \mathscr{C}$
18: *// On vérifie le raffinement de B, i.e. si* R_B *est homogène*
19: **if** $\mathscr{C}(R_B) <= \sigma(R_B)$ **then**
20: $\mathscr{X} \leftarrow \mathscr{X} \cup \{B\}$
21: **else**
22: $\mathscr{X} \leftarrow \mathscr{X} \cup R_B$
23: $\mathscr{P} \leftarrow \mathscr{X}$
return $\mathscr{P}$

Notre algorithme détaillé ci-dessous est composé de trois étapes principales.

— La première étape (lignes 1 et 2 de notre algorithme) consiste tout simplement à extraire les boîtes de texte depuis l'image grâce à l'outil OCR. Le résultat est un ensemble de boîtes de texte.
— La deuxième étape (ligne 3 et 4) consiste à appliquer l'algorithme k-means sur l'ensemble des boîtes de texte. La partition obtenue est notée.$\mathscr{P}$.
— La troisième étape (lignes 7-21) consiste à raffiner, de manière progressive, le regroupement obtenu avec chacun des critères restants. Au préalable, pour chaque critère, l'algorithme k-means est appliqué sur l'ensemble des boîtes de textes donné par l'algorithme OCR. Ensuite, à chaque élément B de regroupement est éclaté (selon la Définition 1) avec le résultat de regroupement des boîtes de texte avec l'algorithme k-means (ligne 10). Si l'entropie associée au résultat de l'éclatement, alors l'élément B est remplacé par son éclatement (lignes 18-21).

Nous utilisons donc cette mesure d'entropie pour évaluer l'homogénéité des différentes classes entre un *cluster* sur un critère donné et un ensemble de *clusters* d'un autre critère.

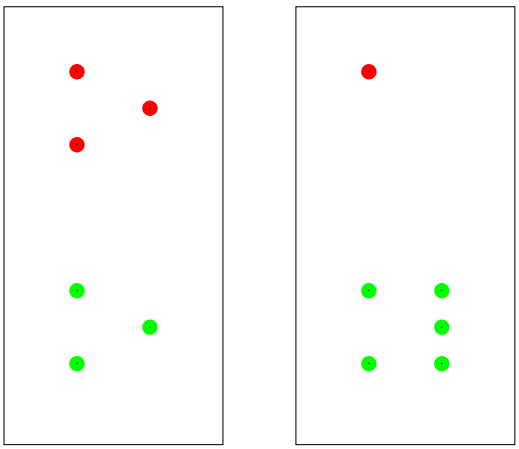

FIGURE 4 – Exemples de deux situations et leurs valeurs d'entropie

Dans l'exemple ci-dessus, chaque point appartient à un même *cluster* pour le critère 1. Les couleurs représentent les différentes classes calculées pour le critère 2.
Pour l'exemple de gauche, qui représente un *cluster* donnée par le critère 1, trois points appartiennent à la classe "rouge" et les trois autres appartiennent à la classe "verte". Les différentes classes du critère 2 au seins de ce *cluster* sont hétérogènes. La valeur d'entropie est maximale (c'est-à-dire de 1).
L'exemple à droite n'a qu'un seul élément appartenant à la classe "rouge". Dans cet exemple, les classes du critère 2 sont beaucoup plus homogènes. En conséquence, la valeur d'entropie de ce *cluster* est plus basse (0,65).

Dans le premier cas, le *cluster* est séparé en deux nouveaux clusters, alors que dans le deuxième cas le *cluster* reste inchangé.

4 Évaluation expérimentale

Les expérimentations conduites dans cet article sont réalisées sur des processeurs `Intel XEON E5-2637 v4` 4 cœurs à 3,6 GHz avec 128 Go de RAM RDIMM, sous `CentOS 7.3 Kernel 3.10`. Concernant la configuration logicielle, nous utilisons la bibliothèque OCR `DocTr version 4.0`[10] et l'implémentation du *clustering* `scikit-learn version 1.3.2` [13].

Nous avons sélectionné 29 images sur lesquelles nous avons exécuté notre procédure de détection de la légende, en utilisant les différents critères présentés plus tôt. Sur chacune de ces cartes, la légende a été étiquetée à la main, afin de comparer le résultat de nos calculs à la "véritable" légende, qui est la réponse optimale. Nous avons pris soin de sélectionner des cartes très différentes les unes des autres, dans leur composition, leur résolution, et la représentation de leur légende. L'ensemble des cartes utilisées dans cette section est disponible à l'adresse `https://www.cril.univ-artois.fr/~marzinkowski/incremental-clustering-results/`

Nous précisons ici que nous comparons des surfaces (ou des zones) de l'image, car ce qui nous intéresse est de savoir si une zone est une légende ou non. L'un des avantages de procéder ainsi, est que si un mot (à l'intérieur d'un groupe de mots) n'est pas détecté par l'OCR, on peut néanmoins récupérer une large partie de la zone (sinon la zone complète si le mot est à l'intérieur d'un groupe de mots) dans notre détection, et ainsi ne pas être trop sensibles aux résultats de l'OCR.

Nous ne reportons pas dans le détail les temps d'exécution, qui sont similaires dans toutes nos expérimentation. Voici toutefois une idée générale du coût calculatoire des différentes étapes de notre algorithme de détection de légende : (i) le temps d'exécution de l'`OCR` sur notre ensemble de données varie entre 10 et 25 secondes, suivant la quantité de textes présents sur la carte donnée en entrée (ii) les cinq ensembles de des partitions obtenues avec k-means sont calculés en environ une seconde (iii) enfin, le temps nécessaire au raffinement n'excède jamais les 100ms. Une synthèse plus précise des temps d'exécution des différentes étapes est donnée dans Table 1.

On observe donc que quelque soit la carte, notre algorithme prend moins de 20 secondes pour effectuer toutes les étapes de la détection.

	Moyenne	Écart type
OCR	17,231	2,095
Clustering	1,036	0,790
Raffinement	0,005	0,002

TABLE 1 – Présentation synthétique des temps d'exécution (en secondes)

4.1 Méthode d'évaluation

Nous avons considéré deux métriques d'évaluation :

— Intersection sur Union (IOU)
— Intersection sur Étiquetage (IOE)

Ces métriques sont calculées en comparant 2 zones rectangulaires de la carte. L'une est la zone calculée par notre algorithme, l'autre est la zone étiquetée manuellement comme réponse optimale.

Sommairement, l'*Intersection sur Union* (IOU), également appelée *indice de Jaccard*, consiste à calculer le ratio entre la surface de l'intersection des 2 zones, et celle de leur union. Pour l'*Intersection sur étiquetage* (IOE), il s'agit du ratio entre l'intersection des deux zones, et la zone étiquetée. Ces deux métriques ont l'intérêt d'être insensible à la résolution des cartes ; l'IOU a par ailleurs été utilisé dans de nombreux travaux [14].

Illustrons ces métriques à l'aide de la Figure 5. Celle-ci contient 3 zones de couleur : rouge, bleu et jaune ; nous noterons respectivement leur aire A_R, A_B et A_J.

On considère que le rectangle étiqueté (réponse optimale) est le rectangle composé des surfaces bleues et jaunes, tandis que la zone calculée par l'algorithme est représentée par les rectangles jaunes et rouges. Ainsi, la surface jaune représente la zone correctement identifiée (vrai positif), la surface bleue la zone indûment non identifiée (faux négatif) et enfin, la surface rouge représente la zone identifiée à tort par l'algorithme comme la zone à détecter (faux positif).

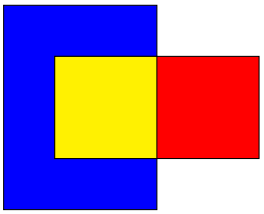

FIGURE 5 – Exemple de différentes surfaces, permettant d'illustrer les métriques IOU et IOE

Dans cet exemple, le score d'IOU est égal à l'aire en jaune divisée par la somme de toutes les aires, soit $(A_J/(A_B + A_R + A_J))$.

Le score d'IOE, quant à lui, est égal à l'aire jaune divisée par la somme des aires bleues et jaunes : $(A_J/(A_B + A_J))$.

Nous avons calculé ces scores pour toutes les permutations des cinq critères pris en compte, ainsi que pour chaque sous-ensemble de critères. Par exemple, si l'on prend en compte l'ordre [Distance, Alignement, Hauteur, Couleur du texte, Couleur de fond], nous calculons un score où les clusters sont calculés avec le seul premier critère (pas de raffinement), un autre avec les deux premiers critères, puis les trois premiers, etc. Ainsi, nous avons été exhaustifs, afin de connaître l'ensemble de critères le plus adapté de manière certaine. Nous obtenons donc pour chaque carte un total de $\sum_{i=1}^{5} \prod_{j=i}^{5} j$ scores, soit 325 scores.

Nous évaluons le résultat en sélectionnant, parmi les *clusters* calculés, ceux dont la métrique d'évaluation est la plus élevée.

4.2 Résultats

Dans la suite de cet article, nous utiliserons les abréviations suivantes pour les 5 critères présentés dans la Section 3.3 :

— D : la distance – équation (5)
— A : l'alignement – équation (3)
— H : la hauteur – équation (7)
— T : la couleur du texte – équation (8)
— F : la couleur de fond – équation (6)

4.2.1 Clustering Mono-Critère

Dans un premier temps, nous avons simplement réalisé un *clustering* en suivant un seul critère, sans aucun raffinement postérieur. Les IOU moyens de ce premier test sont disponibles dans la Table 2.

D	A	H	T	F
42,8%	28,2%	9,8%	14,6%	15,1%

TABLE 2 – IOU moyens sur les 29 cartes d'un *clustering* mono-critère, pour chaque critère

On observe que le critère pourvoyeur des meilleurs résultats est la distance (D). Ce constat n'est pas spécialement surprenant, dans la mesure où une légende regroupe dans la large majorité des cas un ensemble de textes dans une sous-zone spaciale d'une carte ; la proximité de ces textes favorise ce critère de distance.
A l'inverse, le critère le moins favorable -pris individuellement- est la hauteur (H). Encore une fois, ce résultat semble peu surprenant, et peut s'expliquer par le fait que la hauteur n'est pas un élément particulièrement discriminant pour distinguer les éléments textuels d'une légende. Non seulement les textes d'une légende peuvent avoir des tailles différentes (hiérarchie de l'information, etc.), mais d'autres éléments de la carte peuvent avoir la même taille que ces textes particuliers. Il ne semble donc pas pertinent de considérer la hauteur comme seul critère.

4.2.2 De l'importance du raffinement multi-critère

Nous avons poursuivi nos expérimentations en évaluant l'intérêt de combiner nos différents critères via le raffinement successif présenté dans la Section 3.4.
Ici, nous avons voulu montrer l'intérêt du raffinement successif des différents critères, indifféremment de l'ordre dans lequel ceux-ci sont pris. Ainsi, nous avons considéré 5 classes de résultats, suivant le nombre de critères utilisés. Tous les ordres possibles ont été testés, et les résultats suivants indiquent une moyenne de ces résultats.
La Figure 6 se concentre sur l'IOU, et représente le nombre de cartes de notre jeu de données où, en moyenne, un lancement a donné un résultat supérieur à la valeur indiquée en abscisse.
Par exemple, les valeurs pour l'abscisse "10%" indiquent qu'utiliser un seul critère sur les 5 (sans raffinement, comme dans la section précédente) permet d'obtenir en moyenne un IOU supérieur à 10% pour 24 des 29 cartes, tandis qu'utiliser 2,3,4 ou 5 critères porte ce chiffre à 27. Bien sûr, plus les valeurs sur l'axe des abscisses sont élevées, plus le nombre des cartes concernées faiblit.

On note sur cette Figure 6 que le nombre de critères pris en compte a un impact sur la qualité du résultat fourni. En particulier, il est remarquable que chaque raffinement par un nouveau critère permet d'améliorer les résultats. Encore une fois, ceci est vrai indistinctement de l'ordre dans lequel les raffinements sont effectués, puisque ces chiffres indiquent des moyennes de tous les ordres possibles. Il est ici clair qu'exploiter les différentes caractéristiques des éléments textuels de la légende est essentiel à l'efficacité de la procédure de détection.
La Figure 7 reprend la même présentation, mais en s'intéressant cette fois à l'IOE. Assez clairement, les résultats sont ici bien meilleurs. Ceci s'explique aisément par la métrique choisie, IOE étant par construction plus permissive, comme vu dans l'exemple en Figure 5.
Les valeurs élevées représentées par cette Figure illustrent que notre algorithme détecte plutôt correctement les légendes dans de très nombreux cas. Les différences avec l'IOU de la Figure 6 indiquent toutefois que notre technique a tendance à "sur-approximer" le cadre de la légende, cette métrique sanctionnant précisément les approximations trop larges.

4.2.3 Meilleur lancement

Bien que la section précédente fasse abstraction de l'ordre des critères considérés dans les raffinements successifs, celui-ci a bien évidemment une certaine importance dans la performance de notre procédure. Nous nous intéressons ici au lancement qui a obtenu les meilleurs résultats en moyenne.
Conformément à nos observations précédentes, ce lancement utilise l'ensemble des 5 critères pour effectuer sa détection. Intuitivement, on pourrait imaginer que le critère de distance (D), le meilleur dans un *clustering* mono-critère, se trouve en première place dans l'ordre considéré. Or, ce n'est pas le cas. L'ordre optimal que nous avons obtenu est F-A-D-H-T.
C'est donc en regroupant d'abord suivant la couleur de fond (F), puis en raffinant successivement avec l'alignement, la distance, la hauteur et enfin la couleur du texte, que nous obtenons nos meilleurs résultats. Dans cette configuration, l'IOU moyen obtenu sur l'ensemble des cartes est de 52%. Par ailleurs, 7 cartes (sur les 29 de notre jeu de données) obtiennent un IOU supérieur à 80%. l'IOE moyen obtenu pour ce lancement est de 79,48%. Il est à noter que ces scores pourraient paraître faibles, mais nous insistons ici sur le fait que notre jeu de données est hétérogène, et contient des cartes qui sont des défis pour ce type de détection. En particulier, plusieurs cartes ne sont pas conformes aux hypothèses que nous avons faites (tel que l'alignement à gauche).
Nous nous sommes ici concentrés sur un lancement particulier (F-A-D-H-T), cependant, d'autres ordres dans les critères sont également intéressants, et certains *patterns* semblent pouvoir être dessinés. La liste exhaustive des résultats obtenus pendant notre campagne d'expérimentation est disponible à l'adresse `https://www.cril.univ-artois.fr/~marzinkowski/`

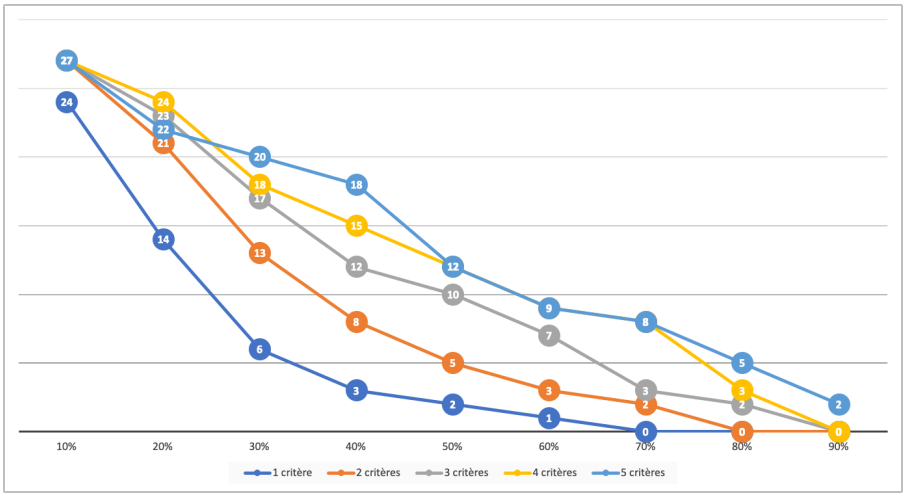

FIGURE 6 – Graphique du pourcentage de carte moyen d'IOU supérieur à l'axes des abscisses

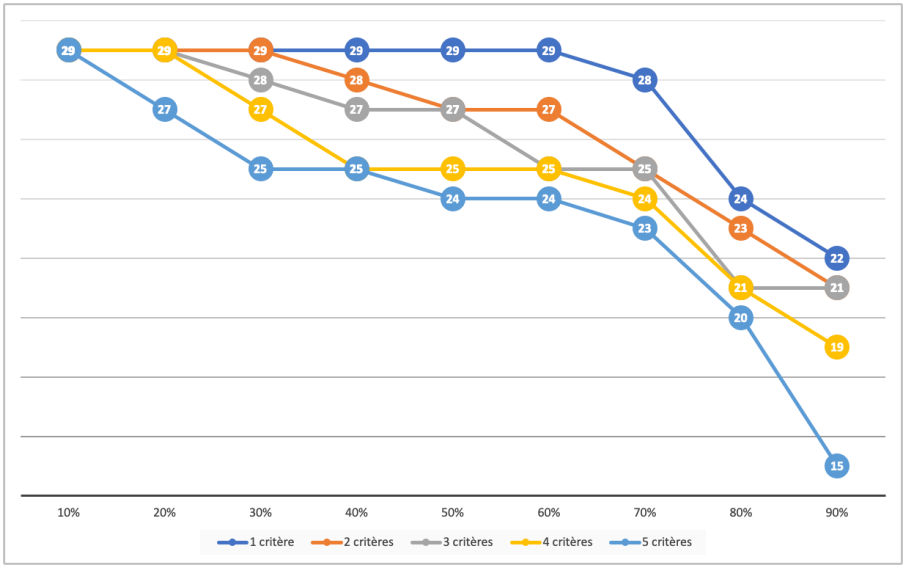

FIGURE 7 – Graphique du pourcentage de carte moyen d'IOE supérieur à l'axes des abscisses

`incremental-clustering-results/`

5 Conclusion et perspectives

Dans cet article, nous avons vu comment les techniques de regroupement, couplées à des critères appropriés, permettent de détecter automatiquement les textes de légende dans des cartes en tout genre : plans de ville, réseaux d'eau, cartes routières, etc.

Pour regrouper efficacement les zones de texte susceptibles de représenter une légende, nous avons établi cinq critères, chacun associé à une mesure de distance spécifique. Par la suite, nous avons introduit un algorithme incrémental, paramétré par un vecteur de critères, nous permettant d'améliorer le partitionnement en exploitant l'entropie comme mesure de l'homogénéité des partitions obtenues à chaque étape. L'étude expérimentale, menée sur un échantillon de cartes très représentatif et utilisant un ou plusieurs critères, a montré que les raffinements successifs des partitions donnent de meilleurs résultats que l'utilisation d'un seul critère.

Il existe plusieurs pistes de recherches futures pour ce travail. Premièrement, nous visons à généraliser les distances proposées pour chaque critère afin de s'adapter aux différentes formes que peuvent prendre les textes de légende. Un autre travail consiste à proposer une fonction d'agrégation globale des différentes distances associées aux cinq critères ; et ainsi appliquer l'algorithme de regroupement une seule fois.

Une autre orientation future consiste à explorer des algorithmes de regroupement alternatifs, avec un accent particulier sur le *clustering* hiérarchique. En parallèle de ces travaux, nous envisageons d'étudier des algorithmes de clustering visant à identifier les objets alignés et associés aux textes de légende. Enfin, nous collectons actuellement un

grand nombre de cartes de légende pour mener une étude expérimentale plus approfondie.

Remerciements

Ce travail a reçu le soutien du programme de recherche Horizon Europe Marie Skłodowska-Curie Actions MSCA (Staff Exchanges) grant agreement 101086252 ; Call : HORIZON-MSCA-2021-SE-01, Projet : STARWARS (STormwAteR and WastewAteR networkS heterogeneous data AI-driven management).
Il a également reçu le soutien de l'Agence Nationale de la Recherche, via le projet ANR CROQUIS (Collecte, représentation, complétion, fusion et interrogation de données de réseaux d'eau urbains hétérogènes et incertaines), grant ANR-21-CE23-0004.

Références

[1] Anurag Agrahari and Rajib Ghosh. Multi-oriented text detection in natural scene images based on the intersection of mser with the locally binarized image. *Procedia Computer Science*, 171 :322–330, 2020.

[2] Yao-Yi Chiang and Craig A Knoblock. An approach for recognizing text labels in raster maps. In *20th International Conference on Pattern Recognition*, pages 3199–3202, 2010.

[3] Youcef Djenouri, Asma Belhadi, Philippe Fournier-Viger, and Jerry Chun-Wei Lin. Fast and effective cluster-based information retrieval using frequent closed itemsets. *Information Sciences*, 453 :154–167, 2018.

[4] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[5] S. Karthikeyan, Vignesh Jagadeesh, and B. S. Manjunath. Learning bottom-up text attention maps for text detection using stroke width transform. In *2013 IEEE International Conference on Image Processing*, pages 3312–3316, 2013.

[6] D. Kavitha and V. Radha. Text detection based on text shape feature analysis with intelligent grouping in natural scene images. In Somnath Bhattacharyya, Jitendra Kumar, and Koeli Ghoshal, editors, *Mathematical Modeling and Computational Tools*, pages 467–479, Singapore, 2020. Springer Singapore.

[7] Rahim Khan, Yurong Qian, and Sajid Naeem. Extractive based text summarization using kmeans and tf-idf. *International Journal of Information Engineering and Electronic Business*, 11 :33–44, 05 2019.

[8] Rutuja Kumbhar, Snehal Mhamane, Harshada Patil, Sukruta Patil, and Shubhangi Kale. Text document clustering using k-means algorithm with dimension reduction techniques. In *5th International Conference on Communication and Electronics Systems (ICCES)*, pages 1222–1228, 2020.

[9] James Lintern. Recognizing text in google street view images. 2010.

[10] Mindee. doctr : Document text recognition. `https://github.com/mindee/doctr`, 2021.

[11] Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. A survey of ocr evaluation tools and metrics. HIP '21. Association for Computing Machinery, 2021.

[12] J. R. Parker. *Algorithms for Image Processing and Computer Vision*. Wiley Publishing, 2nd edition, 2010.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

[14] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union : A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 658–666. Computer Vision Foundation / IEEE, 2019.

[15] Markus Stricker and Markus Orengo. Storage and retrieval for image and video databases (spie) - similarity of color images. *SPIE Proceedings*, 2420 :381–392, March 1995.

[16] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *ICML*, pages 577–584, 2001.

[17] Chucai Yi and YingLi Tian. Text string detection from natural scenes by structure-based partition and grouping. volume 20(9), pages 2594—-2605, 2011.

[18] Hui Yin, Amir Aryani, Stephen Petrie, Aishwarya Nambissan, Aland Astudillo, and Shengyuan Cao. A rapid review of clustering algorithms, 2024.

Techniques neurosymboliques probabilistes pour la classification supervisée informée par la logique

A. Ledaguenel[1,2], C. Hudelot[2], M. Khouadjia[1]

[1] IRT SystemX, Palaiseau
[2] CentraleSupélec, MICS

arthur.ledaguenel@irt-systemx.fr

Résumé

L'IA neurosymbolique est un champ de recherche émergent dont l'objectif est de combiner les capacités d'apprentissage des réseaux de neurones avec les aptitudes de raisonnement des systèmes symboliques. Ce papier présente un formalisme pour la classification supervisée informée par la logique, décrit succinctement les principales tâches et jeux de données traitées par la littérature, puis détaille un ensemble de techniques neurosymboliques basées sur le raisonnement probabiliste et analyse leur complexité asymptotique.

Mots-clés

Neurosymbolique, classification, logique.

Abstract

Neurosymbolic AI is a growing field of research aiming to combine neural networks learning capabilities with the reasoning abilities of symbolic systems. In this paper, we introduce a formalism for supervised multi-label classification informed by propositional background knowledge, desribe the main tasks and datasets tackled in the literature, then present a set of neurosymbolic techniques based on probabilistic logics and analyze their asymptotic complexity.

Keywords

Neurosymbolic, classification, logic.

1 Introduction

L'intelligence artificielle neurosymbolique est un champ de recherche émergent dont l'objectif est de combiner les capacités d'apprentissage des réseaux de neurones avec les aptitudes de raisonnement des systèmes symboliques. Cette hybridation peut prendre de nombreuses formes en fonction de la tâche traitée et des avantages ciblés [22, 44].
Un sous-domaine important de l'IA neurosymbolique est l'apprentissage machine informé (*informed machine learning*) [39], qui étudie comment exploiter de la connaissance *a priori* pour améliorer un système d'apprentissage. De nouveau, les techniques introduites dans la littérature peuvent être de natures très diverses selon le type de tâche (eg. régression, classification, détection, génération, etc.), le formalisme utilisé pour représenter la connaissance (eg. équations mathématiques, graphes de connaissances, logiques, etc.), l'étape à laquelle la connaissance est intégrée (eg. traitement des données, design de l'architecture du réseau de neurones, procédure d'apprentissage, procédure d'inférence, etc.) ou même les avantages attendus de l'hybridation (eg. explicabilité, performance, frugalité, etc.).
Dans ce papier, nous introduisons un formalisme pour la classification multi-classes supervisée informée par la logique.
Les contributions et le plan de l'article sont les suivants. Après quelques notions préliminaires sur la classification neuronale, la logique propositionnelle et le raisonnement probabiliste en Section 2, nous introduisons en Section 3 notre nouveau formalisme pour représenter une tâche de classification multi-classes supervisée informée par la logique. La connaissance *a priori* est exprimée par une formule propositionnelle qui décrit l'ensemble des combinaisons de classes sémantiquement *valides*. Nous illustrons ce formalisme par quelques exemples de tâches et jeux de données les plus fréquemment utilisés dans la littérature neurosymbolique. Puis nous nous appuyons sur ce formalisme dans la Section 4 pour reformuler les principales techniques neurosymboliques probabilistes existantes et analysons leurs principales propriétés dans la Section 5. Nous discutons dans la Section 6 des problèmes de complexité relatifs à ces techniques. Enfin, nous exposons les travaux connexes en Section 7 et concluons en Section 8 avec des pistes de recherche pour de futurs travaux.

2 Préliminaires

2.1 Classification neuronale

En apprentissage machine supervisé, l'objectif est d'apprendre une relation fonctionnelle $f : \mathcal{X} \mapsto \mathcal{Y}$ entre un **domaine d'entrée** $\mathcal{X}$ et un **domaine de sortie** $\mathcal{Y}$ à partir d'un jeu de données annoté $\mathsf{D} := (x^i, \mathbf{y}^i)_{1 \leq i \leq d} \in (\mathcal{X} \times \mathcal{Y})^d$. Les systèmes d'apprentissage profond sont habituellement décrits en deux modules : un réseau de neurones profond (*i.e.* un graphe computationnel paramétrique et différentiable) M est conçu pour émuler au mieux la fonction f et un module de coût différentiable L est utilisé pour mesurer la distance entre les prédictions et les annotations. Les poids du réseau sont alors optimisés en utilisant la descente de gra-

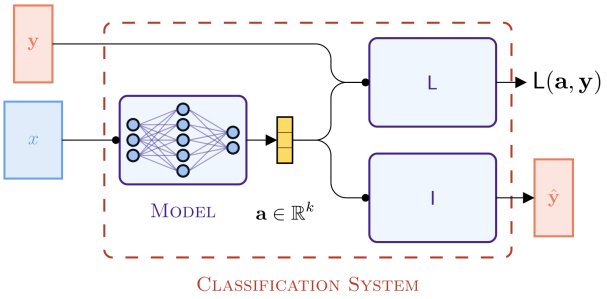

FIGURE 1 – Illustration d'un système neuronal de classification

dient afin de minimiser l'erreur empirique.
Cependant, cette description ne peut pas être appliquée telle qu'elle pour des tâches de classification. La classification multi-classes est un type de tâches d'apprentissage pour lesquelles les éléments de sortie sont des sous-ensemble d'un ensemble fini de classes $\mathbf{Y}$. On appelle un tel sous-ensemble un **état**, et le domaine de sortie (qui contient l'ensemble des états) est noté $\mathcal{Y} = \mathbb{B}^{\mathbf{Y}}$, avec $\mathbb{B} = \{0, 1\}$. Un état $\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}$ peut également être vu comme une fonction qui associe à chaque variable une valeur dans $\mathbb{B}$ (*i.e.* une variable est associée à 1 par l'état si elle appartient au sous-ensemble décrit par l'état). Étant donné que l'espace de sortie est discret, un module de coût différentiable ne peut pas être défini sur $\mathcal{Y}$[2].
Ainsi, nous adoptons une description légèrement modifiée, dans laquelle un troisième module I, appelé le module d'inférence, doit être défini pour combler l'espace entre la nature continue du réseau de neurones (nécessaire à la descente de gradient) et la nature discrète du domaine de sortie. Ce troisième module, bien qu'essentiel, est rarement explicitement défini. Une illustration de cette description d'un système neuronal de classification peut être trouvée sur la Figure 1.

Définition 1. Un **système neuronal de classification** est constitué de :

- un module **paramétrique différentiable** (*i.e.* neuronal) M, appelé le **modèle**, qui prend en entrée une instance $x \in \mathcal{X}$, des paramètres $\theta \in \Theta$ et donne en sortie un vecteur de scores $\mathsf{M}(x, \theta) := \mathsf{M}_\theta(x) := \mathbf{a} \in \mathbb{R}^k$, appelé **scores pré-activation** ou **logits**.
- un module **non-paramétrique différentiable** L, appelé le module de **perte**, qui prend en entrée des logits $\mathbf{a} \in \mathbb{R}^k$ une annotation $\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}$ et donne en sortie un scalaire positif $\mathsf{L}(\mathbf{a}, \mathbf{y})$.
- un module **non-paramétrique** I, appelé le module d'**inférence**, qui prend en entrée des logits $\mathbf{a} \in \mathbb{R}^k$ et donne en sortie une prédiction $\hat{\mathbf{y}} \in \mathbb{B}^{\mathbf{Y}}$.

Une approche classique pour définir un système neuronal de classification découle d'une interprétation probabiliste. Les logits produits par le réseau de neurones sont vus comme des paramètres d'une distribution conditionnelle de probabilité sur le domaine de sortie en fonction de l'entrée $\mathcal{P}(\cdot|\mathsf{M}_\theta(x))$, le module de perte calcule l'entropie croisée de cette distribution avec les annotations tandis que le module d'inférence donne la prédiction la plus probable étant donné la distribution apprise.
Quand aucune connaissance *a priori* n'est disponible (cas non-informé), il est classique de faire l'hypothèse que les variables de sorties sont indépendantes. On illustre ci-dessous comment cette hypothèse peut se traduire dans un système neuronal de classification.

Indépendante Pour la classification indépendante multi-classes (*cim*), on applique une couche *sigmoïde* par dessus le réseau pour transformer les logits en scores de probabilité. Le module de perte calcule l'entropie croisée entre ces scores et l'annotation, tandis que le module d'inférence prédit 1 pour toutes les variables dont la probabilité dépasse 0.5 et 0 pour les autres. Cela se traduit par les modules suivants :

$$\mathsf{L}_{cim}(\mathbf{a}, \mathbf{y}) := \mathtt{BCE}(\mathsf{s}(\mathbf{a}), \mathbf{y})$$
$$= -\sum_j y_j . \log(\mathsf{s}(a_j)) + (1 - y_j) . \log(1 - \mathsf{s}(a_j)) \quad (1)$$
$$\mathsf{I}_{cim}(\mathbf{a}) := \mathbb{1}[\mathbf{a} \geq 0] \quad (2)$$

où $\mathsf{s}(a_i) = \frac{e^{a_i}}{1+e^{a_i}}$ est la fonction sigmoïde, $\mathtt{BCE}$ est l'entropie croisée binaire et $\mathbb{1}[z] := \begin{cases} 1 & \text{si z vrai} \\ 0 & \text{sinon} \end{cases}$ est la fonction indicatrice.

2.2 Logique propositionnelle

Une signature propositionnelle est un ensemble $\mathbf{Y}$ de symboles appelés **variables** (*e.g.* $\mathbf{Y} = \{a, b\}$). Une **formule propositionnelle** est formée de manière inductive en combinant variables et autres formules par des connecteurs unaires ($\neg$, qui exprime la négation) et binaires ($\vee, \wedge$, qui expriment la disjonction et la conjonction respectivement). Par exemple, la formule $\kappa = a \wedge b$ exprime la conjonction des deux variables a et b. Un **état** de $\mathbf{Y}$ est une application $\mathbf{y} : \mathbf{Y} \mapsto \mathbb{B}$, où $\mathbb{B} := \{0, 1\}$. Un état $\mathbf{y}$ peut être prolongé en une **valuation** $\mathbf{y}^*$ sur l'ensemble des formules en suivant la sémantique usuelle (*e.g.* $\mathbf{y}^*(a \wedge b) = \mathbf{y}(a) \times \mathbf{y}(b)$). On dit qu'un état $\mathbf{y}$ **satisfait** une formule κ, noté $\mathbf{y} \models \kappa$, ssi $\mathbf{y}^*(\kappa) = 1$. Une formule est dite **satisfiable** s'il existe un état qui la satisfait. Deux formules sont dites équivalentes, noté $\kappa \equiv \gamma$, ssi elles sont satisfaites par les mêmes états. Dans la suite du papier et sauf mention contraire on supposera que l'ensemble de variables est fini et on notera $\mathbf{Y} = \{Y_j\}_{1 \leq j \leq k}$. On se réfère à [41] pour plus de détails sur la logique propositionnelle.

2.3 Raisonnement probabiliste

Un défi pour l'IA neurosymbolique est de combler l'écart entre la nature discrète de la logique et la nature continue des réseaux de neurones. Dans cette section, on définit les notion de distributions discrètes et de raisonnement probabiliste afin de fournir une interface entre ces deux univers.
Une **distribution de probabilité** sur un ensemble fini de variables $\mathbf{Y}$ est une application $\mathcal{P} : \mathbb{B}^{\mathbf{Y}} \mapsto \mathbb{R}^+$ qui associe un réel positif $\mathcal{P}(\mathbf{y})$ à chaque état $\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}$, de sorte que la somme des valeurs donne $\sum_{\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}} \mathcal{P}(\mathbf{y}) = 1$. Pour pouvoir

définir des opérations internes, comme la multiplication de deux distributions, nous étendons cette définition à des distributions non-normalisées $\mathcal{E} : \mathbb{B}^{\mathbf{Y}} \mapsto \mathbb{R}^+$. La distribution nulle associe tous les états à 0. La fonction de **partition** $\mathsf{Z} : \mathcal{E} \mapsto \sum_{\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}} \mathcal{E}(\mathbf{y})$ donne la somme de la distribution sur tous ses états. On note $\overline{\mathcal{E}} := \frac{\mathcal{E}}{\mathsf{Z}(\mathcal{E})}$ la distribution **normalisée** (lorsque $\mathcal{E}$ est non nulle). Le **mode** d'une distribution $\mathcal{E}$ est son état le plus probable, *i.e.* $\operatorname{argmax}_{\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}} \mathcal{E}(\mathbf{y})$.

Une distribution classique dans la littérature est la distribution exponentielle, qui est au coeur des modules de perte et d'inférence utilisés pour la classification indépendante multi-classes.

Définition 2. Étant donné un vecteur $\mathbf{a} \in \mathbb{R}^k$, on définit la **distribution exponentielle** comme :

$$\mathcal{E}(\cdot|\mathbf{a}) : \mathbf{y} \mapsto \prod_{1 \leq i \leq k} e^{a_i . y_i}$$

On note également $\mathcal{P}(\cdot|\mathbf{a}) = \overline{\mathcal{E}(\cdot|\mathbf{a})}$ la distribution de probabilité correspondante.

Remarque 1. La distribution exponentielle est la distribution jointe de variables indépendantes de Bernouilli $\mathcal{B}(p_i)_{1 \leq i \leq k}$ avec $p_i = \mathsf{s}(a_i)$, où $\mathsf{s}(\mathbf{a}) = (\frac{e^{a_j}}{1+e^{a_j}})_{1 \leq j \leq k}$ est la fonction sigmoid.

Exemple 1. *La Table 1 représente la distribution exponentielle sur deux variables Y_1 et Y_2 paramétrée par le vecteur $\mathbf{a} := (2, -1)$. La fonction de partition est $\mathsf{Z}(\mathcal{E}(\cdot|\mathbf{a})) = 11.5$. Le mode de la distribution $\mathcal{P}(\mathbf{y}|\mathbf{a})$ est $\hat{\mathbf{y}} = (1, 0)$ avec une probabilité de 0.64.*

y_1	y_2	$\mathcal{E}(\mathbf{y}\|\mathbf{a})$	$\mathcal{P}(\mathbf{y}\|\mathbf{a})$
0	0	e^0	0.09
0	1	e^{-1}	0.03
1	0	e^2	**0.64**
1	1	e^1	0.24
		11.5	1

TABLE 1 – Représentation tabulaire d'une distribution.

Traditionnellement, lorsqu'une croyance (*belief*) a propos de variables aléatoires est exprimée par une distribution de probabilité et que de nouvelles informations sont collectées sous la forme d'observations (*evidence*), deux choses nous intéressent : calculer la probabilité de ces observations et mettre à jour nos croyances en utilisant la règle de Bayes, en conditionnant la distribution sur les observations. Le raisonnement probabiliste nous permet d'effectuer les mêmes opérations avec de la connaissance logique à la place d'observations. Prenons une distribution de probabilité $\mathcal{P}$ sur un ensemble de variables $\mathbf{Y} := \{Y_j\}_{1 \leq j \leq k}$ et une formule propositionnelle **satisfiable** κ sur ce même ensemble de variables. Notons $\mathbb{1}_\kappa$ la fonction indicatrice de κ qui associe 1 aux états qui satisfont κ et 0 aux autres :

$$\mathbb{1}_\kappa(\mathbf{y}) = \begin{cases} 1 & \text{si } \mathbf{y} \models \kappa \\ 0 & \text{sinon} \end{cases}$$

Définition 3. La **probabilité** de κ sous $\mathcal{P}$ est la somme des probabilités des états qui satisfont κ, *i.e.* :

$$\mathcal{P}(\kappa) := \mathsf{Z}(\mathcal{P} \cdot \mathbb{1}_\kappa) = \sum_{\mathbf{y} \in \mathbb{B}^{\mathbf{Y}}} \mathcal{P}(\mathbf{y}) \cdot \mathbb{1}_\kappa(\mathbf{y}) \tag{3}$$

La distribution $\mathcal{P}$ **conditionée par** κ, notée $\mathcal{P}(\cdot|\kappa)$, est :

$$\mathcal{P}(\cdot|\kappa) := \overline{\mathcal{P} \cdot \mathbb{1}_\kappa} \tag{4}$$

Remarque 2. Les deux définitions données ci-dessus sont sémantiques et non syntaxiques : elles s'appuient uniquement sur l'ensemble des états qui satisfont la formule et pas sur la syntaxe de la formule, ce qui signifie que deux formules équivalentes auront la même probabilité et la même distribution conditionnée.

Lorsque la distribution utilisée est une distribution de probabilité exponentielle $\mathcal{P}(\cdot|\mathbf{a})$, on note :

$$\mathcal{P}(\kappa|\mathbf{a}) := \mathsf{Z}(\mathcal{P}(\cdot|\mathbf{a}).\mathbb{1}_\kappa) \tag{5}$$

$$\mathcal{P}(\cdot|\mathbf{a}, \kappa) := \frac{\mathcal{P}(\cdot|\mathbf{a}).\mathbb{1}_\kappa}{\mathcal{P}(\kappa|\mathbf{a})} \tag{6}$$

Étant donné que la distribution $\mathcal{P}(\cdot|\mathbf{a})$ est strictement positive (pour tous $\mathbf{a}$), si κ est satisfiable, alors $\mathcal{P}(\kappa|\mathbf{a}) > 0$. Calculer $\mathcal{P}(\kappa|\mathbf{a})$ est un problème de comptage appelé **Probabilistic Query Estimation** (PQE). Calculer le mode de $\mathcal{P}(\cdot|\mathbf{a}, \kappa)$ est un problème d'optimisation appelé **Most Probable Explanation** (MPE). Résoudre ces deux problèmes se révélera au coeur de plusieurs techniques neurosymboliques que l'on introduira (voir Section 4).

Exemple 2. *La Table 2 reprend la distribution de l'Exemple 1 et illustre le raisonnement probabiliste sur la formule $\kappa = \neg Y_1 \vee Y_2$. La probabilité de κ est $\mathcal{P}(\kappa|\mathbf{a}) = 0.36$. Le mode de la distribution $\mathcal{P}(\cdot|\mathbf{a}, \kappa)$ est $\hat{\mathbf{y}} = (1, 1)$ avec une probabilité de 0.67.*

y_1	y_2	$\mathcal{P}(\mathbf{y}\|\mathbf{a})$	$\mathcal{P}(\mathbf{y}\|\mathbf{a}) \cdot \mathbb{1}_\kappa(\mathbf{y})$	$\mathcal{P}(\mathbf{y}\|\mathbf{a}, \kappa)$
0	0	0.09	0.09	0.24
0	1	0.03	0.03	0.09
1	0	0.64	0	0
1	1	0.24	0.24	**0.67**
		1	**0.36**	1

TABLE 2 – Représentation tabulaire d'une distribution.

3 Classification supervisée informée par la logique propositionnelle

On dit qu'une tâche de classification multi-classes supervisée est **informée** lorsque lui est attachée de la connaissance *a priori*, exprimée par une formule propositionnelle κ satisfiable, qui spécifie quels états du domaine de sortie $\mathcal{Y}$ sont **sémantiquement valides**.

Un jeu de données supervisé D est **cohérent** avec la formule κ si toutes les annotations la satisfont (*i.e.* $\forall 1 \leq i \leq$

$n, \mathbf{y}^i \models \kappa$). Dans ce papier, nous faisons l'hypothèse que les jeux de données d'entraînement et de test sont cohérents à la connaissance *a priori*. Cependant, certaines techniques permettent d'assouplir cette hypothèse et de travailler avec des jeux de données contenant des incohérences.
Les techniques évoquées dans ce papier ne s'intéressent pas à l'architecture du réseau en elle-même (eg. perceptron multi-couches, réseau convolutif, réseau récurent, transformer, etc.), qui dépend principalement de la structure du domaine d'entrée (eg. images, textes, etc.), mais se focalisent sur les deux autres modules afin d'intégrer notre connaissance *a priori* sur le domaine de sortie. On donne ci-dessous quelques exemples du type de structures qui peuvent être exprimées en logique propositionnelle, ainsi que de la manière dont on peut intégrer cette connaissance dans les modules de perte et d'inférence.

Catégorique Dans une tâche de classification **catégorique**, une et une seule variable est classifiée comme *vraie* dans chaque état valide. Cette contrainte peut être facilement exprimée en logique propositionnelle :

$$\kappa_{\odot_k} := \left(\bigvee_{1 \leq j \leq k} Y_j \right) \wedge \left(\bigwedge_{1 \leq j < l \leq k} (\neg Y_j \vee \neg Y_l) \right) \quad (7)$$

où la première partie impose qu'une variable soit classifiée comme *vraie* et la seconde partie empêche deux variables d'être *vraies* en même temps.
Pour la classification catégorique, la couche *sigmoid* est habituellement remplacée par une couche *softmax*, et la variable avec le score de probabilité maximum est prédite, ce qui donne les modules suivants :

$$\mathsf{L}_{\odot_k}(\mathbf{a}, \mathbf{y}) := \mathtt{CE}(\mathtt{s}(\mathbf{a}), \mathbf{y}) = -\log(\langle \sigma(\mathbf{a}), \odot_k(j) \rangle) \quad (8)$$

$$\mathsf{I}_{\odot_k}(\mathbf{a}) := \odot_k(\mathrm{argmax}(\mathbf{a})) \quad (9)$$

où CE est l'entropie croisée, $\sigma(\mathbf{a}) = (\frac{e^{a_j}}{\sum_l e^{a_l}})_{1 \leq j \leq k}$ et $\odot_k$ donne le *one-hot encoding* (en commençant à 1) de $j \in [\![1, k]\!]$, *e.g.* $\odot_4(2) = (0, 1, 0, 0)$.

Exemple 3 (MNIST)**.** *MNIST [28] est l'un des jeux de données les plus vieux de la vision par ordinateur et consiste en des petites images de chiffres manuscrits (e.g. 4 ou 7). La tâche de classification catégorique a pour domaine d'entrée l'espace des images* 28×28 *en niveaux de gris, et une classe pour chaque chiffre. Comme décrit plus haut, la connaissance a priori sur la tâche est simplement exprimée par la formule* $\kappa_{\odot_{0:9}}$.

Ce type de tâches peut être élargi pour inclure les tâches **multi-catégoriques** pour lesquelles l'ensemble de variables peut être partitionné en plusieurs groupes de variables catégoriques. Le formule propositionnelle serait alors une conjonction de formule catégoriques sur des ensembles de variables disjoints.

Exemple 4 (Leptograpsus)**.** *Le jeu de données Leptograpsus [9] décrit* 5 *mesures morphologiques effectuées sur* 200 *crabes de couleurs et de sexes différents. L'objectif de la tâche de classification multi-catégorique associée est de prédire la couleur et le sexe d'un crabe à partir de ces mesures. Le domaine d'entrée est l'espace des mesures morphologiques* $\mathcal{X} = \mathbb{R}^5$ *et le domaine de sortie est l'ensemble des états des variables* $\mathbf{Y} = \{\mathsf{r}, \mathsf{b}, \mathsf{f}, \mathsf{m}\}$ *pour les classes* ***red***, ***blue***, ***female*** *et* ***male*** *respectivement. La connaissance a priori sur la tâche impose que chaque crabe soit d'une et une seule couleur, et d'un et un seul sexe, i.e. :*

$$\kappa := (\mathsf{r} \vee \mathsf{b}) \wedge (\neg \mathsf{r} \vee \neg \mathsf{b}) \wedge (\mathsf{f} \vee \mathsf{m}) \wedge (\neg \mathsf{f} \vee \neg \mathsf{m})$$

Hiérarchique La classification **hiérarchique** sur un ensemble de variables $\mathbf{Y}$ est usuellement représentée par un graphe dirigé acyclique $G = (\mathbf{Y}, E_h)$ où les noeuds sont les variables et les arrêtes E_h exprime l'inclusion d'une classe dans une autre (*e.g.* un chien est un animal). Lorsque le graphe est un arbre (ou une forêt) on parle de classification **taxonomique**. Ce formalisme peut également être enrichi par des arrêtes d'exclusion $H = (\mathbf{Y}, E_h, E_e)$ (*e.g.* une instance ne peut pas être classifiée comme chien et chat simultanément), comme dans les HEX-graphs [15]. Là encore, la traduction en logique propositionnelle vient naturellement :

$$\kappa_H := \left(\bigwedge_{(i,j) \in E_h} Y_i \vee \neg Y_j \right) \wedge \left(\bigwedge_{(i,j) \in E_e} (\neg Y_i \vee \neg Y_j) \right) \quad (10)$$

où la première partie s'assure qu'une classe fille ne peut être *vraie* que si sa classe mère l'est également et la seconde partie empêche deux classes mutuellement exclusives d'être *vraies* en même temps.
Dans le cas hiérarchique il n'y a pas de consensus dans la littérature sur les modules de pertes et d'inférence à utiliser : [36] définit un module de perte hiérarchique pour plus pénaliser les erreurs faîtes sur les plus hautes classes de la hiérarchies (en conservant le module d'inférence de *cim*), [21] affine les logits en fonction de la hiérarchie tandis que [15] conditionne la distribution exponentielle par la connaissance hiérarchique.

Exemple 5 (Cifar-100)**.** *Cifar-100 dataset [25] est un jeu de données composé de 60,000 images classifiées en 20 macro-classes (e.g. reptile), chacune divisée en 5 micro-classes (e.g. crocodile, dinosaur, lizard, turtle, et snake). Le domaine d'entrée est l'espace des images RGB de taille* 32×32. *Typiquement utilisé dans un cadre catégorique sur les 100 micro-classes, il peut également être utilisé dans le cadre d'une tâche de classification hiérarchique en incluant les macro-classes. Les relations hiérarchiques d'une classe macro vers ses classes micro, ainsi que les relations d'exclusion entre les macro-classes et entre les micro-classes peuvent être encodées dans un HEX-graphe* $H = (Y, E_h, E_e)$ *et donc exprimées par une formule propositionnelle* κ_H.

Exemple 6 (ImageNet)**.** *Le ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [40] est un challenge de classification d'images, qui a eu lieu annuellement entre 2010 et 2017, et s'est imposé comme un benchmark de référence en vision par ordinateur pour comparer les performances des systèmes neuronaux de classification. En août*

(a) MNIST-Sudoku [4]

(b) Warcraft Shortest Path [3]

FIGURE 2 – Exemples d'instances provenant de jeux de données structurés

2014, le jeu de données ImageNet contenait 14,197,122 images annotées classifiées selon 21,841 synsets de la hiérarchie WordNet [35]. Comme pour Cifar-100 (voire Exemple 5), le jeu de données est le plus souvent utilisé dans un cadre catégorique, mais peut être également utilisé dans un cadre hiérarchique en mobilisant les relations d'inclusion entre synsets définies dans WordNet.

Au delà de ces exemples, la logique propositionnelle peut être utilisée pour définir des domaines de sortie structurés très divers : addition de chiffres manuscrits (*i.e.* k-Add-MNIST) [24], solutions d'un Sudoku [4, 14], chemins dans un graphe dirigé [38, 45, 3], classements [45], matching parfait dans un graphe [38, 1], etc.

4 Techniques neurosymboliques probabilistes

L'objectif d'une technique neurosymbolique est de construire un système neuronal de classification automatiquement à partir de la connaissance *a priori* disponible sur le domaine de sortie, généralisant ainsi les travaux menés sur les cas particuliers de la classification multi-classes indépendante, catégorique et hiérarchique au cas général d'une formule propositionnelle.

Comme précisé plus haut, on ne s'intéresse qu'à la spécification des modules de perte et d'inférence. En particulier, nous étudierons un ensemble de techniques qui s'appuient sur le raisonnement probabiliste pour définir leurs modules. Nous utilisons (abusivement) l'appellation *techniques probabilistes* pour les décrire, même si ces techniques ne disposent pas nécessairement d'une interprétation probabiliste.

Régularisation sémantique Une première technique neurosymbolique utilise une approche multi-objectifs : un terme de régularisation mesurant la cohérence des sorties du modèle avec la connaissance *a priori* est ajouté à l'entropie croisée afin d'optimiser les paramètres sur un double objectif de performance et de cohérence. Il existe plusieurs manières de calculer ce terme de régularisation. Introduit dans un premier temps en utilisant de la logique floue [17, 23, 5], une version basée sur le raisonnement probabiliste est présentée par [45].

Définition 4. Un système neuronal performe la **régularisation sémantique** (*rs*) sur κ ssi ses modules de perte et d'inférence sont :

$$\mathsf{L}^{\lambda}_{rs}(\mathbf{a},\mathbf{y}) = \mathsf{L}_{cim}(\mathbf{a},\mathbf{y}) - \lambda.\log(\mathcal{P}(\kappa|\mathbf{a})) \quad (11)$$

$$\mathsf{I}_{cim}(\mathbf{a}) = \underset{\mathbf{y}\in\mathbb{B}^{\mathbf{Y}}}{\operatorname{argmax}}\mathcal{P}(\mathbf{y}|\mathbf{a}) \quad (12)$$

Conditionnement sémantique Suivant l'interprétation probabiliste introduite en Section 2.1, une manière naturelle d'intégrer de la connaissance *a priori* κ dans un système neuronal de classification est de conditionner la distribution $\mathcal{P}(\cdot|\mathsf{M}_\theta)$ par κ. Ce conditionnement affecte les modules de perte et d'inférence, qui reposent tous deux sur la distribution de probabilité paramétrée par la sortie du réseau de neurones, et conduit à la technique neurosymbolique suivante.

Définition 5. Un système neuronal performe le **conditionnement sémantique** (*cs*) sur κ si ses modules de perte et d'inférence sont :

$$\mathsf{L}_{|\kappa}(\mathbf{a},\mathbf{y}) = -\log(\mathcal{P}(\mathbf{y}|\mathbf{a},\kappa)) \quad (13)$$

$$\mathsf{I}_{|\kappa}(\mathbf{a}) = \underset{\mathbf{y}\in\mathbb{B}^{\mathbf{Y}}}{\operatorname{argmax}}\mathcal{P}(\mathbf{y}|\mathbf{a},\kappa) \quad (14)$$

Les modules de pertes et d'inférence présentés pour la classification indépendante et la classification catégorique sont des cas particuliers du conditionnement sémantique. De même, [15] définit le conditionnement sémantique pour les tâches de classification hiérarchiques. Cette technique est généralisée par [3] à un ensemble de distributions de probabilités définies sur des circuits arithmétiques.

Conditionnement sémantique à l'inférence Le conditionnement sémantique à l'inférence est dérivé du conditionnement sémantique, mais applique le conditionnement uniquement sur le module d'inférence (*i.e.* prédit la sortie la plus probable qui satisfait κ) et conserve le module de perte indépendant.

Définition 6. Un système neuronal performe le **conditionnement sémantique à l'inférence** (*csi*) sur κ ssi ses modules de perte et d'inférence sont :

$$\mathsf{L}_{cim}(\mathbf{a},\mathbf{y}) = -\log(\mathcal{P}(\mathbf{y}|\mathbf{a})) \quad (15)$$

$$\mathsf{I}_{|\kappa}(\mathbf{a}) = \underset{\mathbf{y}\in\mathbb{B}^{\mathbf{Y}}}{\operatorname{argmax}}\mathcal{P}(\mathbf{y}|\mathbf{a},\kappa) \quad (16)$$

En plus de conserver des propriétés clé du conditionnement sémantique (voir Section 5), le conditionnement sémantique à l'inférence se démarque des deux autres techniques en n'intégrant la connaissance que dans le module d'inférence, sans impact sur l'entraînement, ce qui présente deux avantages majeurs. Premièrement, tandis que *cs* et *rs* nécessitent de résoudre PQE pour calculer leur module de perte, *csi* ne nécessite que de résoudre MPE pour son module d'inférence, ce qui est plus facile pour certaines classes de formules (*i.e.* matchings parfaits). Deuxièmement, intégrer la connaissance uniquement à l'inférence offre plus de flexibilité. Par exemple, *csi* peut être utilisé dans le cas

où la connaissance n'est pas disponible pendant l'entraînement. C'est une propriété particulièrement importante à l'ère des **modèles sur étagère** et **modèles de fondations** [8], qui sont pré-entraînés sur des grands volumes de données généralistes avant d'être affinés et combinés sur une multitude de tâches hétérogènes, puisque la connaissance spécifique aux différentes tâches ne peut pas être intégrée pendant la majeure partie de l'entraînement.

5 Propriétés

Nous détaillons ci-dessous quelques propriétés théoriques des techniques présentées. Nous renvoyons le lecteur à [30] pour trouver les définitions et les preuves formelles des propriétés énoncées.

Insensibilité syntaxique Une technique neurosymbolique est **insensible à la syntaxe** (*invariant to syntax*) si deux formules équivalentes aboutissent à des modules de pertes et d'inférence identiques. Toutes les techniques qui s'appuient sur le raisonnement probabiliste sont insensibles à la syntaxe (voir Remarque 2), mais ce n'est pas le cas des techniques qui utilisent la logique floue [5].

Consistance Une technique neurosymbolique est consistante (*consistent*) si les prédictions du module d'inférence satisfont nécessairement la connaissance *a priori* κ. Le conditionnement sémantique ainsi que le conditionnement sémantique à l'inférence sont toutes deux consistantes. En revanche, les techniques neurosymboliques qui n'intègrent la connaissance que dans le module de perte (comme la régularisation sémantique) ne peuvent pas garantir la consistance.

Supérieur à l'inférence Un module d'inférence L_1 est dit supérieur à un module d'inférence L_2 ssi, quel que soit les scores d'entrée du module, si la prédiction de L_2 est consistante avec la connaissance a priori, alors la prédiction de L_1 est identique à celle de L_2. Cette propriété est intéressante car elle garantit, sous l'hypothèse de consistance du jeu de données, que la performance du module L_1 sera nécessairement meilleure (au sens large) que celle du module L_2 si on les évalue sur le même réseau de neurones (avec les mêmes poids entraînés). En particulier, on peut montrer que le module d'inférence conditionné $\mathsf{I}_{|\kappa}$ est supérieur au module d'inférence indépendant I_{cim}. Cette garantie théorique se traduit expérimentalement, comme montré dans [30].

6 Algorithmes et complexité

Après avoir introduit les principales techniques neurosymboliques probabilistes dans la section précédente, nous nous attaquons dans cette section à la question de leur implémentation et de leur complexité. Alors que de nombreux papiers pointent les problèmes de passage à l'échelle des ces techniques, leur complexité asymptotique n'est pas étudiée de manière systématique dans la littérature neurosymbolique. Ceci mène à plusieurs déboires : certaines techniques sont illustrées sur des tâches pour lesquelles le passage à l'échelle n'est pas possible, tandis que certaines tâches tractables sont considérées intractables.

On remarque premièrement que toutes les techniques mentionnées précédemment s'appuient sur la résolution de problèmes de PQE et MPE sur des distributions exponentielles. Ces problèmes sont malheureusement #P-complet et NP-complet (par réduction de #SAT et SAT respectivement) et sont donc **intractables** en général. Il devient alors naturel de se demander pour quelles familles de formules ces problèmes peuvent être résolus en temps polynomial. Nous appelons ces familles des **familles tractables** et étudions certains cas dans la Section 6.2.
Pour ce faire, nous présentons d'abord deux approches de résolution des problèmes de PQE et MPE, basées sur les modèles graphiques et la compilation de connaissance respectivement. On identifie ensuite quelques familles de formules tractables et intractables.

6.1 Algorithmes

Modèles graphiques Les modèles graphiques [27, 43] permettent de spécifier une famille de distributions de probabilité sur un ensemble de variables à travers une représentation graphique. Le graphe encode un ensemble de propriété (*e.g.* factorisation, indépendance, etc.) que les distributions qui appartiennent à la famille doivent respecter. Ces propriétés peuvent alors être exploitées afin de produire une représentation compressée des distributions et d'exécuter des algorithmes d'inférence efficaces [26]. Dans le contexte des logiques propositionnelles probabilistes, le graphe primaire d'une formule κ indique le modèle graphique auquel appartiennent les distributions exponentielles conditionnées par κ. En particulier, les algorithmes classiques de passage de messages sur un arbre de jonction peuvent être utilisés pour résoudre les problèmes de PQE et MPE sur ces distributions, avec une complexité en temps $\mathcal{O}(k2^{\tau(\kappa)})$ où k est le nombre de variables et $\tau(\kappa)$ est la largeur d'arbre du graphe primaire de κ.

Compilation de connaissances La compilation de connaissances [11] consiste à traduire une formule propositionnelle dans un langage de représentation qui permet d'effectuer certaines opérations de manière tractable. Les Sentential Decision Diagrams (SDD) [12] (voir Figure 3) est un langage de représentation qui permet la négation, la disjonction et la conjonction en temps polynomial (en la taille du circuit), ainsi que le calcul de PQE et MPE dans un temps linéaire (en la taille du circuit). De plus, il a été montré qu'une formule κ en forme normale conjonctive de largeur d'arbre $\tau(\kappa)$ peut être traduite dans un *trimmed and compressed* SDD de taille $\mathcal{O}(k2^{\tau(\kappa)})$ [12]. En raison de ces propriétés, les SDD sont devenus un langage standard de représentation des connaissances pour les systèmes neurosymboliques probabilistes [45, 3].

Programmation Linéaire Binaire Comme souligné dans [37], une formule propositionnelle en forme normale conjonctive peut être facilement compilée dans un programme linéaire équivalent (*i.e.* qui est satisfait par les mêmes états). Ainsi, la tâche de MPE sur cette formule peut alors être résolue en utilisant les nombreux algorithmes combinatoires qui ont été développés pour les problèmes de programmation linéaire binaire ou mixte [7].

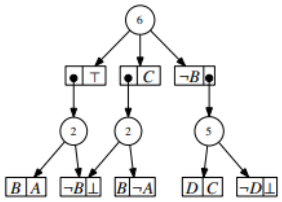

FIGURE 3 – Représentation graphique d'un SDD dans [12]

6.2 Familles tractables

Une familles de formules est dite **tractable** si et seulement si pour toute formule κ de la famille il est possible d'effectuer PQE et MPE sur n'importe quelle distribution exponentielle conditionnée par κ en un temps polynomial (en le nombre de variables de la formule).

Largeur d'arbre bornée Comme énoncé ci-dessus, une condition suffisante de tractabilité d'une famille de formules (et commune aux modèles graphiques et à la compilation de connaissances) est la possibilité de construire en temps polynomial (dans la taille de la signature) une décomposition en arbre de largeur bornée pour toutes les formules de la famille. Ceci implique directement que la famille des formules **taxonomiques** est tractable, puisque de largeur d'arbre 1. Le cas de k-Add-MNIST est intéressant à cet égard : longtemps jugé intractable par la communauté [24], il s'est avéré qu'une représentation légèrement différent du problème (incluant les retenues successives des additions de chiffres) aboutit à une famille de formules de largeur d'arbre bornée (avec une décomposition constructible en temps linéaire) [32].

Énumérable en temps polynomial Un autre type de famille tractable est celui des familles énumérables en temps polynomial, *i.e*. dont on peut énumérer les solutions de chaque formule dans un temps polynomial (en le nombre de variables de la formule). Il est facile de voir pourquoi MPE et PQE sont tous deux calculables en temps polynomial pour une famille énumérable : il suffit de lister chaque état valide et sa probabilité pour compter la probabilité de la formule et trouver son état le plus probable. Cette famille recouvre par exemple le cas des formules **catégoriques** et celui des formules **hiérarchiques** sur un **HEX-graphe en forme d'arbre** et **saturé**.

Cependant, ces deux conditions n'épuisent pas l'ensemble des familles tractable, comme le montre les exemples suivants.

Formules multi-catégoriques La famille des formules multi-catégoriques est un bon exemple des limites des critères de largeur d'arbre et d'énumérabilité pris séparément. En effet, il est facile de voir qu'il s'agit d'une famille de largeur d'arbre non bornée (la largeur d'arbre d'une formule catégorique est son nombre de variables) et non énumérable en temps polynomial (le nombre d'états valides est potentiellement exponentiel). Cependant, il est possible de décomposer chaque formule en composantes indépendantes et énumérables, ce qui rend la famille tractable.

Chemins simples La famille des formules qui encodent les chemins (simples) dans un graphe dirigé acyclique est elle aussi de largeur d'arbre non bornée et non énumérable. Cependant, en suivant un ordre topologique des arrêtes du graphe, il est possible de compiler cette formule en un OBDD [11] de taille polynomiale (dans le nombre d'arêtes) [29]. Cette propriété permet d'assurer la tractabilité de MPE et PQE pour cette famille. Il est de plus intéressant de remarquer que le calcul de MPE pour cette famille peut se ramener facilement à un calcul de plus court chemin, et donc être résolu avec des algorithmes combinatoires classiques (*e.g*. Bellman-Ford [6] ou Dijkstra [16]) : les probabilités sur les arrêtes sont transformés en scores réels par une sigmoïde inverse et multipliés par -1. Cette équivalence avec le problème de plus court chemin nous indique également qu'une famille de formules qui encodent les chemins simples dans un graphe dirigé (potentiellement cyclique) est **intractable**. En effet, le calcul de plus court chemin (avec poids négatifs) est un problème NP-complet, par réduction du problème d'existence d'un cycle Hamiltonien [19]. De plus, le calcul de PQE est #P-complet pour cette famille, par réduction au problème de comptage des chemins simples dans un graphe dirigé [42].

7 Travaux connexes

Logiques alternatives Il existe d'autres langages et sémantiques que la logique propositionnelle pour exprimer de la connaissance *a priori* sur une tâche de classification. Si certains correspondent à un fragment de la logique propositionnelle, comme les HEX-graphs dans [15], d'autres lui sont incommensurables, comme la programmation logique avec sémantique des modèles stables dans [46] ou la programmation par contraintes linéaires dans [37], voire passent à l'ordre supérieur, comme le langage de programmation logique Prolog [13] dans [33] ou la logique de premier ordre dans [5]. Les compromis pour ces différents langages sont principalement entre leur concision, leur expressivité et leur tractabilité. Les conséquences du choix de langage en termes de complexité de calculs ne sont pas encore bien comprises.

Logiques floues De nombreux travaux utilise les **logiques floues** [20, 23, 5] à la place du raisonnement probabiliste comme moyen de faire le pont entre la nature discrète de la connaissance et la nature continue du réseau de neurones.

Logiques pondérées Dans notre formalisme, nous avons fait l'hypothèse que la connaissance *a priori* est une contrainte dure, systématiquement satisfaite dans les jeux d'entraînement et de test. Certains formalismes permettent un langage plus souple qui permettent de représenter de l'incertitude sur la connaissance *a priori*. Des logiques pondérées permettent par exemple d'exprimer quelles formules ont *le plus de chances* d'être satisfaites et peuvent être intégrées dans des systèmes neurosymboliques [10, 34]. Les travaux de la *theory of evidence* [31] ou des *probability kinematics* [18] offrent des outils théoriques qui permettent de combiner de l'information provenant de deux sources incertaines (*e.g*. un réseau de neurones et de la connaissance

a priori probabiliste).

Apprentissage non supervisé De nombreux travaux explorent le potentiel de l'IA neurosymbolique dans un cadre d'apprentissage non *pleinement* supervisé (*i.e.* chaque instance du jeu de données est annotée sur l'ensemble des classes), comme l'apprentissage **faiblement supervisé** (*i.e.* certaines instances ne pas annotées sur toutes les classes) ou **semi-supervisé** (*i.e.* certaines instances ne sont pas annotées). Les techniques de régularisation [2, 45, 23, 5, 17] en particulier sont particulièrement indiquées pour l'apprentissage semi-supervisé et montrent que l'intégration de connaissances *a priori* dans le processus d'apprentissage peut grandement diminuer le besoin en instances annotées tout en augmentant les performances du système. Les techniques de conditionnement sémantique [3, 33, 46] permettent de traiter certaines variables comme des variables latentes en marginalisant la distribution apprise, et ne nécessitent donc pas de labels sur les variables latentes pendant l'apprentissage.

8 Conclusion

Après avoir présenté un formalisme pour la classification supervisée informée par la logique propositionnelle, ce papier réalise une revue de littérature des jeux de données structurés ainsi que des techniques neurosymboliques probabilistes. Enfin, nous détaillons quelques résultats relatifs à la complexité asymptotique des techniques probabilistes, qui manque d'une étude systématique dans la littérature.
Les pistes de recherche pour nos futurs travaux incluent, entre autres : une meilleure cartographie des familles tractables pour les différentes techniques probabilistes, l'utilisation de logiques alternatives, une étude des cadres d'apprentissages non-supervisés en IA neurosymbolique, ou encore l'apprentissage simultané du modèle neuronal et de la structure de la tâche.

Remerciements

Ce travail a été soutenu par le gouvernement français dans le cadre du programme "France 2O30" au sein de l'Institut de Recherche Technologique SystemX.

Références

[1] Kareem AHMED, Kai-Wei CHANG et Guy VAN DEN BROECK. « Semantic Strengthening of Neuro-Symbolic Learning ». In : *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Sous la dir. de Francisco RUIZ, Jennifer DY et Jan-Willem van de MEENT. T. 206. Proceedings of Machine Learning Research. PMLR, 25–27 Apr 2023, p. 10252-10261. URL : https://proceedings.mlr.press/v206/ahmed23a.html.

[2] Kareem AHMED et al. « Neuro-symbolic entropy regularization ». In : *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*. Sous la dir. de James CUSSENS et Kun ZHANG. T. 180. Proceedings of Machine Learning Research. PMLR, jan. 2022, p. 43-53. URL : https://proceedings.mlr.press/v180/ahmed22a.html.

[3] Kareem AHMED et al. « Semantic Probabilistic Layers for Neuro-Symbolic Learning ». In : *Advances in Neural Information Processing Systems*. Sous la dir. de S KOYEJO et al. T. 35. Curran Associates, Inc., 2022, p. 29944-29959.

[4] Eriq AUGUSTINE et al. « Visual Sudoku Puzzle Classification : A Suite of Collective Neuro-Symbolic Tasks ». In : *International Workshop on Neural-Symbolic Learning and Reasoning*. 2022.

[5] Samy BADREDDINE et al. « Logic Tensor Networks ». In : *Artificial Intelligence* 303 (2022), p. 103649. ISSN : 0004-3702. DOI : https://doi.org/10.1016/j.artint.2021.103649. URL : https://www.sciencedirect.com/science/article/pii/S0004370221002009.

[6] Richard BELLMAN. « On a Routing Problem ». In : *Quarterly of Applied Mathematics* 16.1 (1958). Publisher : Brown University, p. 87-90. ISSN : 0033-569X. URL : https://www.jstor.org/stable/43634538 (visité le 23/04/2024).

[7] M. BENICHOU et al. « Experiments in mixed-integer linear programming ». In : *Mathematical Programming* 1.1 (1er déc. 1971), p. 76-94. ISSN : 1436-4646. DOI : 10.1007/BF01584074. URL : https://doi.org/10.1007/BF01584074 (visité le 02/05/2024).

[8] Rishi BOMMASANI et al. « On the Opportunities and Risks of Foundation Models ». In : *CoRR* abs/2108.07258 (2021). arXiv : 2108.07258. URL : https://arxiv.org/abs/2108.07258.

[9] N. A. CAMPBELL et R. J. MAHON. « A multivariate study of variation in two species of rock crab of the genus Leptograpsus ». In : *Australian Journal of Zoology* 22.3 (1974). Publisher : CSIRO PUBLISHING, p. 417-425. ISSN : 1446-5698. DOI : 10.1071/zo9740417. URL : https://www.publish.csiro.au/zo/zo9740417 (visité le 02/05/2024).

[10] Alessandro DANIELE et Luciano SERAFINI. « Knowledge Enhanced Neural Networks ». In : *PRICAI 2019 : Trends in Artificial Intelligence : 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part I*. Berlin, Heidelberg : Springer-Verlag, 26 août 2019, p. 542-554. ISBN : 978-3-030-29907-1. DOI : 10.1007/978-3-030-29908-8_43. URL : https://doi.org/10.1007/978-3-030-29908-8_43 (visité le 19/04/2024).

[11] Adnan DARWICHE. *Modeling and Reasoning with Bayesian Networks*. Cambridge : Cambridge University Press, 2009. ISBN : 978-0-521-88438-9. DOI : 10.1017/CBO9780511811357. URL : https://www.cambridge.org/core/books/modeling-and-reasoning-with-bayesian-networks/8A3769B81540EA93B525C4C2700C9DE6 (visité le 07/08/2023).

[12] Adnan DARWICHE. « SDD : A New Canonical Representation of Propositional Knowledge Bases ». In : *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. 2011.

[13] Luc DE RAEDT, Angelika KIMMIG et Hannu TOIVONEN. « ProbLog : a probabilistic prolog and its application in link discovery ». In : *Proceedings of the 20th international joint conference on Artifical intelligence*. IJCAI'07. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., jan. 2007, p. 2468-2473. (Visité le 07/08/2023).

[14] Marianne DEFRESNE, Sophie BARBE et Thomas SCHIEX. « Scalable coupling of deep learning with logical reasoning ». In : *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. IJCAI '23. <conf-loc>, <city>Macao</city>, <country>P.R.China</country>, </conf-loc>, 19 août 2023, p. 3615-3623. ISBN : 978-1-956792-03-4. DOI : 10.24963/ijcai.2023/402. URL : https://doi.org/10.24963/ijcai.2023/402 (visité le 23/02/2024).

[15] Jia DENG et al. « Large-Scale Object Classification Using Label Relation Graphs ». In : *Computer Vision – ECCV 2014*. Sous la dir. de David FLEET et al. Springer International Publishing, 2014, p. 48-64. ISBN : 978-3-319-10590-1.

[16] E. W. DIJKSTRA. « A note on two problems in connexion with graphs ». In : *Numerische Mathematik* 1.1 (1er déc. 1959), p. 269-271. ISSN : 0945-3245. DOI : 10.1007/BF01386390. URL : https://doi.org/10.1007/BF01386390 (visité le 21/02/2024).

[17] Michelangelo DILIGENTI, Marco GORI et Claudio SACCÀ. « Semantic-based regularization for learning and inference ». In : *Artificial Intelligence* 244 (mars 2017), p. 143-165. ISSN : 00043702. DOI : 10.1016/j.artint.2015.08.011.

[18] Zoltan DOMOTOR, Mario ZANOTTI et Henson GRAVES. « Probability Kinematics ». In : *Synthese* 44.3 (1980). Publisher : Springer, p. 421-442. ISSN : 0039-7857. URL : https://www.jstor.org/stable/20115538 (visité le 02/05/2024).

[19] Michael R. GAREY et David S. JOHNSON. *Computers and Intractability : A Guide to the Theory of NP-Completeness*. USA : W. H. Freeman & Co., 1979. 338 p. ISBN : 978-0-7167-1044-8.

[20] Francesco GIANNINI et al. « T-norms driven loss functions for machine learning ». In : *Applied Intelligence* 53.15 (fév. 2023), p. 18775-18789. ISSN : 0924-669X. DOI : 10.1007/s10489-022-04383-6. URL : https://doi.org/10.1007/s10489-022-04383-6 (visité le 07/08/2023).

[21] Eleonora GIUNCHIGLIA et Thomas LUKASIEWICZ. « Coherent Hierarchical Multi-Label Classification Networks ». In : *Advances in Neural Information Processing Systems*. T. 33. Curran Associates, Inc., 2020, p. 9662-9673. URL : https://proceedings.neurips.cc/paper/2020/hash/6dd4e10e3296fa63738371ec0d5df818-Abstract.html (visité le 13/09/2023).

[22] Henry A. KAUTZ. « The Third AI Summer : AAAI Robert S. Engelmore Memorial Lecture ». In : *AI Mag.* 43 (2022), p. 93-104.

[23] Emile van KRIEKEN, Erman ACAR et Frank van HARMELEN. « Analyzing Differentiable Fuzzy Logic Operators ». en. In : *Artificial Intelligence* 302 (jan. 2022), p. 103602. ISSN : 0004-3702. DOI : 10.1016/j.artint.2021.103602. URL : https://www.sciencedirect.com/science/article/pii/S0004370221001533 (visité le 07/08/2023).

[24] Emile van KRIEKEN et al. « A-NeSI : A Scalable Approximate Method for Probabilistic Neurosymbolic Inference ». In : *Advances in Neural Information Processing Systems* 36 (13 fév. 2024). URL : https://proceedings.neurips.cc/paper_files/paper/2023/hash/4d9944ab3330fe6af8efb9260aa9f307-Abstract-Conference.html (visité le 21/02/2024).

[25] Alex KRIZHEVSKY. *Learning Multiple Layers of Features from Tiny Images*. 2009.

[26] Frank R. KSCHISCHANG, Brendan J. FREY et Hans Andrea LOELIGER. « Factor graphs and the sum-product algorithm ». In : *IEEE Transactions on Information Theory* 47.2 (2001), p. 498-519. ISSN : 00189448. DOI : 10.1109/18.910572. (Visité le 28/03/2022).

[27] Steffen L. LAURITZEN. *Graphical Models*. Clarendon Press, 1996. ISBN : 978-0-19-852219-5.

[28] Yann LECUN et al. « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86 (11 1998), p. 2278-2323. ISSN : 00189219. DOI : 10.1109/5.726791.

[29] Arthur LEDAGUENEL, Céline HUDELOT et Mostepha KHOUADJIA. *Complexity of Probabilistic Reasoning for Neurosymbolic Classification Techniques*. 2024. arXiv : 2404.08404 [cs.AI].

[30] Arthur LEDAGUENEL, Céline HUDELOT et Mostepha KHOUADJIA. *Improving Neural-based Classification with Logical Background Knowledge*. 2024. arXiv : 2402.13019 [cs.AI].

[31] Jianbing MA et al. « Bridging jeffrey's rule, agm revision and dempster conditioning in the theory of evidence ». In : *International Journal on Artificial Intelligence Tools* 20.4 (août 2011). Publisher : World Scientific Publishing Co., p. 691-720. ISSN : 0218-2130. DOI : 10.1142/S0218213011000401. URL : https://www.worldscientific.com/doi/10.1142/S0218213011000401 (visité le 02/05/2024).

[32] Jaron MAENE et Luc DE RAEDT. « Soft-Unification in Deep Probabilistic Logic ». In : *Advances in Neural Information Processing Systems* 36 (15 déc. 2023). URL : https://papers.nips.cc/paper_files/paper/2023/hash/bf215fa7fe70a38c5e967e59c44a99d0-Abstract-Conference.html (visité le 21/02/2024).

[33] Robin MANHAEVE et al. « Neural probabilistic logic programming in DeepProbLog ». en. In : *Artificial Intelligence* 298 (sept. 2021), p. 103504. ISSN : 0004-3702. DOI : 10.1016/j.artint.2021.103504. URL : https://www.sciencedirect.com/science/article/pii/S0004370221000552 (visité le 07/08/2023).

[34] Giuseppe MARRA et Ondřej KUŽELKA. « Neural markov logic networks ». In : *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. Sous la dir. de Cassio de CAMPOS et Marloes H. MAATHUIS. T. 161. Proceedings of Machine Learning Research. PMLR, 27–30 Jul 2021, p. 908-917. URL : https://proceedings.mlr.press/v161/marra21a.html.

[35] George A. MILLER. « WordNet ». In : *Communications of the ACM* 38 (11 nov. 1995), p. 39-41. ISSN : 15577317. DOI : 10.1145/219717.219748. URL : https://dl.acm.org/doi/10.1145/219717.219748.

[36] Bruce R. MULLER et W. SMITH. « A Hierarchical Loss for Semantic Segmentation ». In : *VISIGRAPP*. 2020. URL : https://api.semanticscholar.org/CorpusID:215791996.

[37] Mathias NIEPERT, Pasquale MINERVINI et Luca FRANCESCHI. « Implicit MLE : Backpropagating Through Discrete Exponential Family Distributions ». In : *Advances in Neural Information Processing Systems*. T. 34. Curran Associates, Inc., 2021, p. 14567-14579. URL : https://proceedings.neurips.cc/paper_files/paper/2021/hash/7a430339c10c642c4b2251756fd1b484-Abstract.html (visité le 17/01/2024).

[38] Marin Vlastelica POGANČIĆ et al. « Differentiation of Blackbox Combinatorial Solvers ». en. In : sept. 2019. URL : https://openreview.net/forum?id=BkevoJSYPB (visité le 27/10/2023).

[39] Laura von RUEDEN et al. « Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems ». In : *IEEE Transactions on Knowledge and Data Engineering* 35.1 (jan. 2023). Conference Name : IEEE Transactions on Knowledge and Data Engineering, p. 614-633. ISSN : 1558-2191. DOI : 10.1109/TKDE.2021.3079836.

[40] Olga RUSSAKOVSKY et al. « ImageNet Large Scale Visual Recognition Challenge ». In : *International Journal of Computer Vision* 115 (3 déc. 2015), p. 211-252. ISSN : 15731405. DOI : 10.1007/s11263-015-0816-y.

[41] Stuart RUSSELL et Peter NORVIG. *Artificial Intelligence A Modern Approach (4th Edition)*. Pearson Higher Ed, 2021. Chap. 7, p. 208-250.

[42] Leslie G. VALIANT. « The Complexity of Enumeration and Reliability Problems ». In : *SIAM Journal on Computing* 8.3 (1er août 1979), p. 410-421. ISSN : 0097-5397. DOI : 10.1137/0208032. URL : https://doi.org/10.1137/0208032 (visité le 21/02/2024).

[43] M J WAINWRIGHT et al. « Graphical Models, Exponential Families, and Variational Inference ». In : *Foundations and Trends R in Machine Learning* 1.2 (2008), p. 1-305. DOI : 10.1561/2200000001. (Visité le 07/04/2022).

[44] Wenguan WANG, Yi YANG et Fei WU. *Towards Data-and Knowledge-Driven Artificial Intelligence : A Survey on Neuro-Symbolic Computing*. 2023. arXiv : 2210.15889 [cs.AI]. URL : https://arxiv.org/abs/2210.15889.

[45] Jingyi XU et al. « A Semantic Loss Function for Deep Learning with Symbolic Knowledge ». In : *35th International Conference on Machine Learning, ICML 2018*. T. 12. International Machine Learning Society (IMLS), 2018, p. 8752-8760. ISBN : 9781510867963.

[46] Zhun YANG, Adam ISHAY et Joohyung LEE. « NeurASP : Embracing Neural Networks into Answer Set Programming ». In : *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. California : International Joint Conferences on Artificial Intelligence Organization, juill. 2020, p. 1755-1762. ISBN : 978-0-9992411-6-5. DOI : 10.24963/ijcai.2020/243.

Conférences Invitées

Généralisation et réseaux de neurones profonds, le cas du TAL et de la RI

E. Gaussier[1]

[1] LIG - CNRS, Université Grenoble Alpes

Eric.Gaussier@imag.fr

Résumé

Les réseaux de neurones profonds, comme les modèles de langue préentraînés sur de grandes collections textuelles, représentent à l'heure actuelle le paradigme dominant en traitement automatique des langues (TAL) et en recherche d'information (RI). Ceci étant, il y a toujours de nombreuses questions quant à leur performance et à leur fonctionnement. En particulier, s'ils ont conduit à des améliorations significatives dans presque toutes les tâches de TAL et de RI, plusieurs études ont mis en avant leurs limites en termes de généralisation, liées à leur difficulté à traiter correctement de nouvelles collections ou de nouvelles tâches. Nous étudierons ces limites dans notre présentation et discuterons les pistes envisagées pour les dépasser.

Mots-clés

Traitement Automatique des Langues, Recherche d'information, Réseaux de Neurones profonds.

Abstract

Deep neural networks, like language models pre-trained on large text collections, currently represent the dominant paradigm in Natural Language Processing (NLP) and information retrieval (IR). That said, there are still many questions surrounding their performance and operation. In particular, while they have led to significant improvements in almost all NLP and IR tasks, several studies have highlighted their limitations in terms of generalizability, linked to their difficulty in correctly handling new collections or tasks. We will examine these limitations in our presentation and discuss the avenues being considered to overcome them.

Keywords

Natural Language Processing, Information Retrieval, Deep Neural Networks.

..

Curriculum Vitae

Éric Gaussier est professeur à l'Université Grenoble Alpes et membre du Laboratoire d'informatique de Grenoble (LIG - CNRS/Université Grenoble Alpes). Il cherche à faire définir de nouveaux réseaux de neurones profonds adaptés au traitement automatique des langues et à la recherche d'information. Il souhaite y intégrer des propriétés inspirées des capacités humaines dès la construction des modèles, afin de les rendre aussi performants sur de nouveaux corpus que sur ceux sur lesquels ils ont été entraînés. Sa nomination en tant que membre senior au titre de la chaire innovation de l'Institut universitaire de France lui permet d'approfondir ses travaux.

Grands modèles de langue : l'avenir du traitement automatique des langues en santé ?

P. Zweigenbaum[1],

[1] LISN, CNRS, Université Paris-Saclay

pz@lisn.fr

Résumé

Le traitement automatique des langues est sous le feu des projecteurs grâce à la notoriété récente des grands modèles de langue. Je tenterai de cerner les points clés de ce succès, de montrer comment nous en sommes arrivés là, et soulignerai les difficultés qui restent à résoudre. J'examinerai dans ce contexte le rôle que les grands modèles de langue peuvent jouer dans le domaine médical, les enjeux qu'ils soulèvent dans ce domaine, et comment arriver néanmoins à en bénéficier, notamment dans le contexte français.

Mots-clés

Traitement Automatique des Langues, grands modèles de , IA et Santé.

Abstract

Automatic language processing is in the spotlight thanks to the recent notoriety of Large Language Models (LLMs). I will attempt to identify the key points of this success, show how we got here, and highlight the difficulties that remain to be solved. In this context, I'll examine the role that large language models can play in the medical field, the issues they raise in this area, and how we can nevertheless manage to benefit from them, particularly in the French context.

Keywords

Natural Language Processing, Large Language Models, AI and Health.

..

Curriculum Vitae

Pierre Zweigenbaum est Directeur de Recherche au CNRS au sein du Laboratoire Interdisciplinaire des Sciences du Numérique (LISN) à l'Université Paris-Saclay, où il a notamment dirigé l'équipe de traitement automatique des langues (ILES). Avant cela, il a passé vingt ans à l'Assistance publique - Hôpitaux de Paris et à l'Inserm. Ses recherches portent sur le traitement automatique des langues appliqué aux textes médicaux. Ses travaux lui ont valu la reconnaissance de l'American College of Medical Informatics (Fellow ACMI, 2014) et de l'International Academy of Health Sciences Informatics (Fellow IAHSI, 2019). Il est également membre du Comité scientifique consultatif de la Plateforme des données de santé. Actuellement coordinateur des projets ANR KEEPHA et PREDHIC sur l'extraction d'informations dans des textes médicaux, il fait partie d'une large collaboration visant à promouvoir des communs numériques pour le traitement automatique du français en santé.