Université Jean Monnet
Saint-Étienne

Habilitation à Diriger des Recherches
de l'Université Jean Monnet

DSPT 9 : Sciences et technologies de l'information et de la communication

Soutenue publiquement le 05/02/2024, par :

Rémi Emonet

Avancées en apprentissage statistique pour la détection d'anomalies et le transfert

Devant le jury composé de :

Rapporteurs :	Marianne Clausel	Professeure	Université de Lorraine
	Nicolas Courty	Professeur	Université de Bretagne Sud
	Paulo Goncalves	DR Inria	ENS Lyon
Examinatrices :	Élisa Fromont	Professeure	Université de Rennes
	Christine Solnon	Professeure	INSA de Lyon
Tuteur :	Marc Sebban	Professeur	Université Jean Monnet

préparée au laboratoire Hubert Curien
UMR 5516 CNRS / Université de Saint-Étienne / Institut d'Optique Graduate School

Table des matières

partie A

Introduction et CV synthétique

Chapitre I

Introduction, trajectoire scientifique et activités d'enseignement

Ce chapitre présente brièvement la structure du manuscrit. Il décrit ma trajectoire scientifique depuis ma thèse et les grands domaines que j'ai abordés. Il résume également mes activités d'enseignement, indissociables de la recherche dans un métier d'enseignant-chercheur.

I.1 Structure et objectifs du manuscrit

exhaustivité raisonnée

Il est difficile de définir le niveau de détail pour un manuscrit censé couvrir l'ensemble de mon activité de recherche mais aussi une partie de celle de mes collaborateurs sur une durée de plus de dix ans. Ce document vise à concilier deux objectifs a priori opposés : celui d'être (quasiment) exhaustif sur les activités menées et celui de rester concis. Dans ce but, certains domaines de recherche seront présentés uniquement sous forme de synthèse en essayant de se concentrer sur une description du contexte, des défis et des contributions ; le lecteur est invité à se référer aux articles pour les détails, en cas d'intérêt particulier pour un thème.

utilité et accessibilité

Certains sujets seront développés plus en détails, parfois en revenant à des éléments fondamentaux d'un domaine mais en supposant que certains autres sont connus du lecteur. Le but visé est de rendre la lecture intéressante, tout en la gardant suffisamment concise mais accessible. Sans illusion sur le fait d'avoir totalement atteint ce but, j'espère rendre la lecture de ce manuscrit utile.

attention obsolescence

Comme certains travaux sont relativement anciens et dans des domaines qui sont souvent restés très actifs, ce document ne fait que, au mieux, présenter les travaux dans leur contexte d'alors. En particulier, la rédaction de ce

manuscrit n'a en aucun cas mené à la ré-évaluation de toutes les approches discutées, ni à une remise à jour de la présentation de l'art.

Tous les articles étant en anglais, il m'a semblé plus utile d'écrire ce manuscrit d'HDR en français. Certains termes seront restés en anglais quand ils sont vraiment trop inhabituels une fois traduits ou que le contexte le justifie (par exemple, un intitulé de cours en anglais). Une version avec l'ensemble des termes traduits peut être mise à disposition au besoin.

légende, guide de lecture, , marges , ,

Le symbole sert à indiquer les parties décrivant directement une contribution. Les autres parties sont soit des éléments d'explication de contexte, soit de réflexions et discussions hors contribution. Des notes de marge sont ajoutées pour permettre de re-parcourir les sections plus efficacement. Le symbole matérialise les liens vers un endroit précis d'un article, pour pouvoir retrouver, en contexte, la partie de celui-ci en lien avec ce manuscrit. Le symbole pointe les astuces qui ne sont pas forcément exposées dans les articles mais qui semblent importantes. Chaque chapitre de contribution commence par entre autre par un trombinoscope dans lequel les personnes (co-)encadrées sont en vert et les co-encadrants en rouge.

PDF et/ou HTML

Ce document existe en version HTML et en deux versions PDF (recto simple/écran/tablette, et recto-verso). Toutes les versions contiennent des liens internes et externes, dans le cas où elles seraient consultées numériquement. Les versions PDF sont en particulier destinées à l'impression recto-verso. De ce fait, tant que possible, des codes QR sont ajoutés pour accéder aux liens. La version HTML est plus facilement navigable et configurable au travers d'un menu (jaune, en haut à droite).

I.2 Diversité des thèmes de recherche

génie logiciel, analyse de vidéo, apprentissage statistique

En préambule, il est à noter que mon parcours fait apparaître deux réorientations thématiques, qui je pense font une des particularités de mon activité. La première réorientation fut majeure, lorsque je suis passé d'une thèse en génie logiciel pour les environnements intelligents à un postdoc en vision par ordinateur et analyse de vidéo. Je dois énormément à Jean-Marc Odobez qui s'est laissé convaincre de mes capacités d'adaptation et qui m'a recruté en postdoc à l'Idiap. La seconde réorientation s'est opérée lors de mon recrutement à l'UJM. À ce moment là, malgré une thèse dans un domaine totalement différent et un postdoc qui n'était pas exactement en lien avec la théorie de l'apprentissage automatique, l'équipe a su me faire confiance pour me permettre de m'adapter à nouveau et a vu mon profil comme une op-

portunité. J'entrerai dans les détails de mes activités de recherche lors de la présentation de mes contributions scientifiques, dans les chapitres suivants.

variété de collaborations et de sujets

Dès mon recrutement en tant que maître de conférences et de part mon profil, il est à noter que j'ai été amené à développer de nombreuses collaborations multidisciplinaires au travers de divers projets. J'ai surtout eu l'opportunité de participer à l'encadrement de plusieurs thèses.

I.3 Activités liées à l'enseignement et la formation

service conséquent, lauréat IUF 2023

Bien que n'ayant eu qu'une année de responsabilité de M1, depuis mon recrutement mon service a été conséquent (240h EqTD en moyenne sur les 5 dernières années) et inclut la gestion d'UE (unité d'enseignement) à fort volume, en particulier en L1 et L2, dans lesquels je suis en charge des CM (cours magistraux) et orchestre l'ensemble du cours incluant les supports, la gestion des intervenants (jusqu'à 9 groupes de TP), l'évaluation des examens. Étant lauréat IUF (Institut Universitaire de France) depuis la rentrée 2023, ma charge d'enseignement est en conséquence considérablement réduite et une partie de cette section doit se lire au passé. J'ai gardé principalement deux cours de master dans lesquels j'étais le moins remplaçable : un cours de programmation web avancée et un cours de modèles probabilistes.

I.3.1 Importance du lien entre formation et recherche

Bien qu'elle ne soit pas directement de la recherche, l'activité d'enseignement est une part importante du quotidien de maître de conférences et dans une certaine mesure, je pense que ces deux activités s'entre-nourrissent. En particulier, je vois deux éléments importants dans ce lien enseignement/recherche.

enseignement pour la recherche

En premier lieu, il est clair que créer, préparer et donner des cours sur des sujets avancés, proches de nos domaines de recherche, nous amènent à approfondir et appréhender ces sujets de manières différentes et à améliorer la compréhension que nous avons de ceux-ci et l'exposition que nous en faisons. Cependant, dans mon cas, rétrospectivement, cet aspect a été assez limité : mon profil varié m'a amené à donner beaucoup de cours dans les domaines de la programmation, du réseau, du génie logiciel, du web. Si ces sujets ne sont pas ceux de mes recherches, ils ont contribué à me rendre plus efficace au quotidien (développement d'outils et visualisations à base de technologies web) et dans les projets (supervision et orientation de choix techniques pour le développement). Si les cours de spécialité n'ont représenté qu'une petite partie de mon service d'enseignement (cours de modèles probabilistes (M2), de machine learning et d'introduction au deep learning), ils m'ont néanmoins permis d'approfondir certains concepts utiles pour ma recherche.

enseignement comme formation doctorale

D'autre part, de nombreux doctorants sont intervenus en TP (travaux pratiques) et/ou TD (travaux dirigés) sur des cours dont j'ai la responsabilité. Au-delà de faire découvrir une des facettes du métier et de permettre au doctorant de se faire une idée et d'acquérir une expérience de l'enseignement, c'est aussi l'occasion d'insister et de pratiquer les aspects pédagogiques, nécessaires dans le contexte de la recherche que ce soit lors de présentations, lors de l'écriture d'articles ou tout simplement lors de l'interaction avec des collègues. De ce fait, l'accompagnement des doctorants dans leur pratique de l'enseignement participe aussi à leur formation à la recherche.

I.3.2 Service d'enseignement

variété de thèmes d'enseignements

De part ma formation initiale, ma thèse avec monitorat incluant des ateliers de pédagogie, mon postdoc et mes recherches courantes, mon profil est assez large et me permet de faire des cours allant de la programmation web avancée aux modèles probabilistes de machine learning, tout en dispensant des cours d'introduction à la programmation (où la pédagogie prime !). Mon postdoc m'a donné l'opportunité d'intervenir sur des cours de niveau doctoral à l'EPFL notamment sur la **perception par ordinateur et les modèles probabilistes**.

liste des enseignements principaux

Je présente ci-dessous une vision synthétique des enseignements que j'ai réalisés depuis mon recrutement (hors intervention ponctuelle) avec pour chaque cours le volume horaire et estimation du nombre d'étudiants.

- **Outils Informatique** (python) (L1 Math/Info/Physique/Chimie, 21-30hEqTD, ~200étu), cours que j'ai monté et gère depuis 2016. Le cours (ré)introduit les bases de la programmation impératives (variables, boucles, listes, fonctions) avec le langage de python. Bien que ne représentant que 14 heures de présentiel, il y a un travail d'organisation très important, avec la recherche et l'orchestration des intervenants, la coordination pédagogique et la gestion des examens.
- **Projet info/méca** (python) (L1 Physique/Chimie, 18-25hEqTD, ~70étu), cours-projet que nous avons monté avec un collègue physicien pour mettre en application les compétences informatiques dans le contexte d'un projet de mécanique du point.
- **Calcul numérique et visualisation** (numpy, matplotlib) (L2 Chimie, ~40hEqTD, ~80étu), cours que j'ai monté pour les non-informaticiens pour qu'ils apprennent les outils de bases qui peuvent leur servir dans leurs domaines d'application.
- **Programmation Orientée Objet** (L3 Info, 36hEqTD, ~45étu), cours que j'ai repris uniquement 1 an (puis rendu) pour compenser l'absence ponctuelle d'un collègue.

- **Développement Web II** (L3 Info, 22-27EqTD, ~45étu), partie client d'un cours, que je fais depuis mon arrivée et qui évolue chaque année. Le cours couvre javascript, l'API fetch, les websockets.
- **Programmation Web Avancée** (M1 Info DSC, 35hEqTD, ~25étu), cours que je fais depuis mon arrivée et qui évolue chaque année. Actuellement, il introduit l'utilisation de Spring, l'écriture d'API Rest et l'utilisation de vuejs.
- **Machine Learning Fundamental and Algorithms** (M1 Info DSC+MLDM, 24hEqTD, ~60étu) que j'ai dispensé pendant deux ans, suite à quoi nous avons restructuré la maquette et les intervenants. Le cours couvrait les arbres de décision, les random forest (bagging) et les réseaux de neurones simples (SLP, MLP) et leur apprentissage.
- **Introduction to Computer Network** (M1 Info DSC+MLDM, 39hEqTD, ~60étu), que j'ai donné depuis mon arrivée, d'abord en français, puis porté en anglais, puis réduit de volume sur ces 5 dernières années et finalement enlevé cette année. Le cours couvrait les fondamentaux du réseau (des principes de routage aux algorithmes utilisés, etc.).
- **Deep Learning** (M1 Info MLDM, 25hEqTD, ~30étu), que j'ai monté l'année passée puis transmis cette année. Ce cours porte sur les bases du deep learning (chain rule, optimisation, losses, architectures, convolutions, etc.).
- **Probabilistic Graphical Models** (M2 Info MLDM, 25hEqTD, ~20étu), que j'ai monté et dispense depuis 3 ans. Ce cours couvre les réseaux bayésiens, les méthodes MCMC, l'inférence variationnelle, les VAE etc.

plateforme d'entraînement, générateur d'examens

Par ailleurs, j'ai développé un générateur d'exercices (plateforme web) qui permet aux étudiants de pratiquer une infinité de variantes de questions types. Ceci est relativement modeste mais j'ai eu de bons retours d'étudiants. Une version est disponible en ligne (la plupart du temps, selon les aléas liés au maintien à jour d'un service), utilisateur hdr, mot de passe « HDR23 ». J'ai également développé un générateur d'examens pratiques ou chaque étudiant reçoit un sujet différent mais sur le même modèle. Ce travail est né de plusieurs constats. D'un côté, un examen pratique s'impose pour les UE d'informatique (et ce déjà avant chatGPT) : à moins de changer les sujets de TP chaque année (et encore), noter les rendus de TP perdait tout son intérêt (tant formatif que sommatif) de part la quantité de rendus copiés ou venant des corrections des années précédentes. D'un autre côté, pour éviter les triches évidentes, comme par exemple une personne rendant le travail de quelqu'un d'autre pendant un examen pratique, il était nécessaire de varier les sujets, à la manière de ce qui peut être fait dans les questionnaires à choix multiples, mais en allant au-delà du mélange de questions et de réponses. J'ai amélioré puis partagé avec les collègues ces outils et envisage de les partager plus largement une fois rendus un peu plus génériques.

Chapitre II

CV : projets, responsabilités, encadrements

II.1 CV Synthétique

Identité

Rémi EMONET – 08/12/1982 – MCF CN – https://home.heeere.com
Université Jean Monnet – Laboratoire Hubert Curien, UMR CNRS 5516

Formation, Diplômes

2024–2029	Lauréat Institut Universitaire de France (IUF) Junior
2005–2009	Thèse en architecture logicielle pour les environnements intelligents (Université de Grenoble)
2004–2005	Master2R Image, Vision, Robotique (Grenoble INP)
2002–2005	Diplôme d'ingénieur de l'Ensimag, mention « Très Bien » (Grenoble)

Expériences Professionnelle (de recherche)

2013–	Maître de Conférences à l'Université Jean Monnet de Saint-Étienne et au Laboratoire Hubert Curien (LabHC), UMR CNRS 5516, thématique « Data Intelligence », équipe « Machine Learning ». Membre de l'équipe-projet Inria MALICE depuis le 01/12/2023.
2010–2013	Post doctorat à l'institut de recherche Idiap, à Martigny, Suisse, dans le cadre du projet européen FP7 VANAHEIM. Intervention en formation doctorale à l'EPFL.
2009–2010	Post doctorat de 6 mois sur les projets CASPER (aide au maintien à domicile) puis MINImage (vision par ordinateur, embarquée au sein des caméras).
2005-2009	Thèse en architecture logicielle pour l'intelligence ambiante dans l'équipe PRIMA de l'INRIA (allocation+monitorat) puis un ATER à l'Ensimag.
2005–2005	Stage M2R, Équipe PRIMA/INRIA, puis contrat dans le projet EU IST CAVIAR.

II.2 Implication dans des projets financés

Depuis 2013 (prise de poste MCF), j'ai déposé ou participé au dépôt de plusieurs projets dont les suivants ont été acceptés :

- **Projet région TADALoT** (*porteur*) (budget UJM 121k€), sur l'apprentissage par transfert et les données déséquilibrées, finançant la thèse de Tanguy Kerdoncuff et une thèse au laboratoire CREATIS.
- **Projet EUR Sleight PIMALEA** (*co-porteur* avec Jean-Philippe Colombier) (budget UJM 89k€), sur l'apprentissage par transfert pour l'interaction laser-matière, finançant le postdoc d'Eduardo Brandao.
- **Projet FUI MIVAO** (*coordinateur local*) (budget UJM 292k€), porté par l'entreprise Bluecime, une thèse (équipe image) et un postdoc (mon équipe).
- **Projet AXA BabyCry** (*coordinateur local LabHC*) (budget UJM 997k€) (début 2024), multidisciplinaire ENES (CNRL) et service néonatalité (CHU), collecte, analyse et apprentissage pour les pleurs de bébés, financement (entre autres) de 2 postdocs dont un en apprentissage automatique.
- **Projet ANR TAUDoS** (en cours, porté par Rémi Eyraud) (budget UJM 182k€), co-encadrement de Volodimir Mitarchuk, distillation de réseaux récurrents (modèles de langage).
- **Projet ANR ROIi** (en cours, coord. LabHC Thierry Fournel), multidisciplinaire IHRIM (UMR 5317, UJM), analyse d'ornements dans les livres imprimés, co-encadrement de Sayan Chaki.
- **Projet Euripides FA4.0** (porté par Olivier Alata, équipe Image) *Failure Analysis*, co-encadrement de la thèse de Leila Jamshidian sur la détection d'anomalies dans des images.
- **Projet ANR APRIORI** (portée par Emilie Morvant) (budget UJM 149k€). J'apporte mon expérience en Deep Learning et approches bayésiennes.
- **Projet ANR Lives** (portée par Élisa Fromont) (budget UJM 142k€), développement des activités en multi-vues dans les travaux de thèse de Valentina Zantedeschi.
- **Projet ANR SoLSTiCe** (portée par Élisa Fromont) (budget UJM 148k€), co-encadrement des thèses de Damien Fourure et de Valentina Zantedeschi.

La liste précédente n'inclue pas les financements moins importants comme des financements de stages. J'ai aussi, durant mon postdoc, travaillé sur deux projets européens, dont principalement le projet FP7 VANAHEIM qui finançait mes travaux sur la décomposition non-supervisée de documents temporels.

II.3 Encadrement de la recherche

De part ma formation, j'ai acquis une expertise multi-domaines qui m'a permis d'être impliqué dans des co-encadrements sur des sujets à large spectre. Je co-encadre actuellement 4 thèses et j'ai co-encadré 7 thèses soutenues.

- **Damien Fourure** (*30% de l'encadrement*) [1]–[5], financé par une bourse doctorale du ministère, sur la segmentation sémantique et l'apprentissage profond pour l'image. Suite à sa thèse, il a été recruté comme senior Data Scientist à EURA NOVA, Belgique. 2017-12-12 (soutenance)
- **Carlos Arango** (*30%*) [6]–[9], financé par une bourse doctorale du ministère, sur la détection de micro-expressions dans les vidéos, suite à un postdoc, il est actuellement chercheur chez Sherpa Engineering. 2018-12-06
- **Valentina Zantedeschi** (*50%*) [10]–[17], financée par le projet ANR Solstice, qui a donné une variété de très bon travaux sur les thèmes de l'apprentissage de métriques et de l'apprentissage statistique en général. Nous avons eu l'occasion de collaborer à nouveau, plus récemment, autour des approches PAC-bayésiennes. Valentina travaille désormais à *ServiceNow Research*, Montréal, Canada. 2018-12-18
- **Kevin Bascol** (*50%*) [18]–[22], thèse CIFRE avec l'entreprise Bluecime, sur la sécurité des télésièges (détection de situations dangereuses à partir d'images par deep learning, avec adaptation de domaine). Suite à sa thèse, il a continué pour 6 mois en postdoc au laboratoire sur le projet FUI MIVAO. Il a ensuite signé un CDI avec son employeur de thèse en tant qu'expert deep learning. 2019-12-16
- **Yichang Wang** (*25%*) [23]–[25], co-encadré avec 3 collègues de Rennes, sur la classification interprétable de séries temporelles. Suite à sa thèse, il a pour l'instant refusé les offres de postes d'enseignant/chercheur en Chine et travaille actuellement chez Total Digital Factory. 2021-09-20
- **Tanguy Kerdoncuff** (*50%*) [26]–[30], financé par le projet région TADALoT que je portais, sur le transport optimal pour l'adaptation de domaine. Il est maintenant chercheur chez Ericsson. 2021-12-09
- **Eduardo Brandao** (*33%*) [31]–[33], sur l'apprentissage automatique guidé par la physique. Il poursuit en postdoc dans l'équipe sur un financement EUR Manutech SLEIGHT (co-porté avec un collègue physicien). 2023-12-20
- **Leila Jamshidian** (*50%*) (*césure* en cours pour des raisons de santé), financée par le projet européen FA4.0, sur la détection d'anomalies dans des images par méthodes auto-supervisées avec gestion d'incertitude. (2021-)
- **Volodimir Mitarchuk** (*25%*), financé sur le projet ANR TAUDoS, sur les aspects théoriques et la distillation de réseaux de neurones récurrents pour l'interprétabilité des modèles de séquences. 2021-
- **Sayan Chaki** (*50%*), financé par le projet pluridisciplinaire ANR ROIi, sur les modèles d'apprentissage de détection non-supervisée d'objets, en lien avec les modèles probabilistes et les auto-encodeurs structurés. 2021-

2022- • **Robin Mermillod-Blondin** (*33%*), en thèse CIFRE avec l'entreprise HID, sur l'optimisation incrémentale multicritère de couleurs plasmoniques pour les documents d'identité.

2022- • **Fayad Ali Banna** (*40%*), financé par un projet IADoc@UDL, à l'interface entre physique et apprentissage automatique, sur l'apprentissage profond pour prédire la formation de structures périodiques de surface induite par laser.

À noter que j'ai collaboré avec différents autres doctorants pendant leur thèse même si je ne les encadrais pas officiellement, notamment **Guillaume Metzler** (soutenance : 2019-09-25), **Rémi Viola** (soutenance : 2022-06-24) et **Paul Viallard** (soutenance : 2022-12-07). J'ai aussi participé informellement à l'encadrement de **Jagannadan Varadarajan** (soutenance : 2012-08-30), pendant ma période de postdoc. J'ai aussi encadré ou co-encadré une vingtaine de stagiaires de M1/M2, ainsi qu'un post-doctorant sur 2 ans (projet MIVAO).

II.4 Collaborations internationales

Depuis le début de ma carrière, j'ai pu entretenir des collaborations internationales avec plusieurs partenaires privilégiés, sur les thématiques suivantes :

- Sur les approches PAC-bayésiennes, avec Pascal Germain (Univ. Laval) et Benjamin Guedj (Inria & University College London) et maintenante Valentina Zantedeschi.
- Sur les modèles liés à la propagation des maladies (depuis 2012), avec entre autres Katayoun Farrahi (Univ. Southampton) et Manuel Cebrían (Univ. Madrid), et plus récemment Petter Holme (Univ. Aalto), Talal Rahwan (New York University Abu Dhabi).
- Via le projet Euripides FA4.0, avec les partenaires du consortium sur la détection d'anomalies dans les images. Cependant, les interactions ont été limitées de part la taille du projet et la situation sanitaire liée au COVID-19 et coïncidant avec le début du projet.
- Sur la bioacoustique, grâce à un contact initié via l'ENES (équipe de bioacoustique de l'université) : Paulo Fonseca (Univ. Lisbonne).
- Pendant mon postdoc, avec notamment Elisabeth Oberzaucher (Univ. Vienne) sur l'analyse éthologique d'activité humaine dans les vidéos, mais aussi avec Thales Italia.
- Ponctuellement
  - Via Software Carpentry, publications avec divers co-auteurs, à l'initiative de Greg V. Wilson (maintenant, Deep Genomics, Toronto).
  - Sur l'optimisation distribuée, avec Jesus Cerquides et Juan Antonio Rodriguez Aguilar (Spanish National Research Concil).
  - Sur l'extraction de cadence de pas dans les vidéos avec Carina Westling et Harry Witchel (UK).

II.5 Responsabilités scientifiques, pédagogiques et d'intérêt collectif

- Depuis 2017, je suis responsable de l'équipe-projet « Machine Learning » du laboratoire Hubert Curien composée de 8 permanents (enseignants-chercheurs) et 10 à 15 non-permanents (doctorants et postdocs). Ceci implique entre autres l'animation d'équipe, la gestion du budget (missions/achats) et les aspects reporting (bilans annuels, évaluations internes, etc).
- Depuis 2015, je suis élu (et ré-élu) à la commission recherche et au conseil académique de l'Université Jean-Monnet. Ceci implique des activités variées, incluant des aspects légaux et administratifs (validation des dossiers de vacataires), d'évaluations des demandes de promotions, de politique générale de l'établissement et de sa stratégie de recherche (équilibre entre disciplines, appels à projet, science ouverte, etc.).
- J'ai été, en 2016, responsable par intérim du M1 DSC (Données et Systèmes connectées).
- Je suis élu au Conseil du laboratoire Hubert Curien depuis 2021, et membre du groupe de travail sur le développement durable de l'UMR. Nous avons notamment été amenés à faire le bilan carbone du laboratoire et à organiser des actions de sensibilisation.

II.6 Expertise et rayonnement

- Participation aux comités de programmes de différentes conférences et relecteur pour des revues internationales de haut niveau, par exemple, CVPR, ICML, ECCV, TPAMI, DAMI, MLJ.
- Contribution à l'animation locale (équipe de recherche et département) via des collaborations et des présentations de type introduction/tutoriel plusieurs fois par an.
- Participation à l'organisation des conférences CAp (2014) et IDA (2015) à Saint-Étienne, ainsi qu'à deux éditions de l'école de printemps « DeepImaging » organisée par le laboratoire Creatis.
- Rapporteur à l'international, pour la thèse de Manuel Vieira qui a utilisé et développé des approches d'apprentissage automatique pour la bioacoustique, domaine lié à une collaboration en cours avec l'ENES.

Au-delà des présentations en interne ou dans le contexte de projets, j'ai exposé mes travaux sur invitation ou dans des journées de tutoriaux :

- Présentation (et participation à l'organisation) d'une école d'été : "DeepImaging" Summer School 2021, deep learning for medical imaging (Labex PRIMES, CNRS).
- Présentation aux journées "Défi IA" sur les approches de détection d'anomalies (2020).
- Présentation au séminaire organisé par l'équipe LACODAM de l'Irisa (2019), sur les limites en haute dimension des approches à base de

distance.

- Présentation à NAVER Labs (ex-Xerox XRCE) sur les travaux de l'équipe (détection d'anomalies, classification déséquilibrée) (2019).
- Présentation (et participation à l'organisation) d'une école d'été : "DeepImaging" Summer School 2019, deep learning for medical imaging (Labex PRIMES, CNRS).
- Présentation en école d'été : Summer School on Transfer Learning (organisée par Worldline) (2018).
- Présentation (et participation à l'organisation) d'un atelier : "Sainté-lyon deep-learning workshop", à Saint-Étienne (2017),
- Présentation aux journées "Optimisation et Machine Learning" (2016).
- Présentation (et participation à l'organisation) d'un atelier : "Sainté-lyon deep-learning workshop", à Lyon (2015),
- Présentation à Xerox XRCE (2015) sur l'adaptation de domaine et les réseaux profonds.
- Présentation au séminaire SMiLe2014.

À noter que je diffuse la plupart de mes présentations en ligne (sauf si très préliminaires ou confidentielles, ou retard) sur la page « research » ou « teach » de mon site), sous des licences permettant leur réutilisation libre.

II.7 Formations et diffusions autres

- Je me suis impliqué dans « Software Carpentry », plutôt au début de ma prise de poste. J'ai notamment donné des formations au CERN et à l'université de Pise.
- J'ai donné des formations « git » au CIRAD de Montpellier, pour le compte du service formation continue de l'université.
- J'ai réalisé les démonstrations sur l'activité « Machine Learning » pour la visite du laboratoire par le jury de l'HCÉRES. Ces démos ont été ré-utilisées pour présenter les travaux de manière ludique et interactive. Quelques vidéos muettes sont disponibles sur mon site.

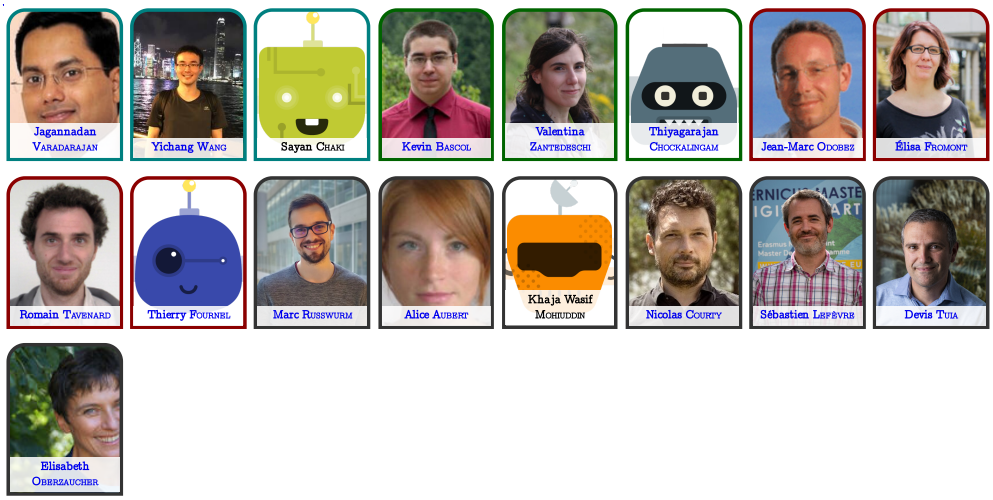
- J'ai co-organisé Django-Girls Saint-Étienne, que nous avons hélas dû annuler à la dernière minute pour cause de vague de COVID-19.
- J'ai participé à des réunions de développeurs et designers (Saint-Étienne ayant été labellisé « French Tech Design Tech ») et j'y suis intervenu pour présenter l'apprentissage automatique. Si l'objectif initial n'était pas lié à la recherche, ceci a néanmoins abouti au recrutement en thèse de Valentina Zantedeschi.

partie B

Travaux réalisés (post thèse)

Chapitre III

Apprentissage non-supervisé et décomposition de documents temporels



Publications: [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] [44] [45] [46] [47] [48] [49] [50] [51] [18] [52] [53] [23] [24] [25] [54].

Projets : VANAHEIM, ROIi.

Codes et liens divers

- Site web relatif aux travaux de postdoc, avec explications et accès au programmes et articles.
  https://www.idiap.ch/resource/research/probamod-v1/
- Site web pour l'article CVPR 2011 [36].
  https://home.heeere.com/publi-2011-cvpr.html
- Supplementary material pour l'article CVPR 2011 [36].
  https://www.idiap.ch/paper/2053/
- Code de la ré-implémentation en python du model HDLSM par Romain Tavenard.
  https://github.com/twitwi/hdlsm

- Code pour les travaux à base d'auto-encodeurs.
  https://github.com/twitwi/Unsupervised-Interpretable-Pattern-Discovery-in-Time-Series-Using-Autoencoders

Présentations

- Unsupervised Decomposition of Images into Motifs
  https://dl.heeere.com/2022-06-16-export-fouille-motifs.pdf
- A Tour of Probabilistic and Deep Approaches for Unsupervised Learning
  http://twitwi.github.io/Presentation-2016-12-13-Journee-Ml-Opt/
- Infinite Mixture Models with Dirichlet Process
  https://github.com/twitwi/Presentation-2015-dirichlet-processes/
- Temporal Topic Models for Probabilistic Motif Mining
  http://www.slideshare.net/remiemonet/temporal-topic-models-for-probabilistic-motif-mining-smile2014

---

interprétation de documents temporels

Le développement de cet axe de recherche a démarré avec mon postdoc à l'Idiap et est encore actif aujourd'hui dans une forme un peu différente. Partant d'un verrou qui était l'interprétation automatique de scène à partir de vidéos (typiquement des vidéos de carrefour avec des voitures), nous avons proposé une formulation plus générale se basant sur des documents temporels. Les données sont sous la forme d'un tableau 2D avec un axe temporel et un axe de features, typiquement une quantité de mouvement à une position donnée dans l'image. Les approches proposées en postdoc sont convolutionnelles selon l'axe temporel. En rendant la formulation convolutionnelle selon les deux axes, cela revient alors à travailler sur la décomposition d'images qui s'apparente ainsi au domaine de la détection d'objet, mais non-supervisée.

prob. concret ⇒ avancées méthodologiques

Comme souvent dans mes activités, les travaux présentés ici partent d'une problématique très concrète. Ceci a donné lieu à une formulation du problème de manière à le rendre plus générique et à pouvoir apporter des solutions plus structurées et plus généralisables.

III.1 Contexte : détection d'anomalies (supervisée et) non-supervisée

anomalies dans les vidéos

Le problème d'origine de mon postdoc était l'interprétation automatique et non supervisée de scène à partir de vidéos, dans un objectif de détection d'anomalies. Plus précisément, il s'agissait, à partir de quelques heures de vidéos prises d'une caméra fixe, d'apprendre à détecter automatiquement les

données anormales. La détection d'anomalies contient en fait deux grandes familles de problèmes.

détection supervisée

Le premier cas est celui de la détection **supervisée** d'anomalies dans laquelle le principal problème est le déséquilibre durant l'apprentissage entre la classe normale et la classe anormale : les anomalies sont beaucoup moins fréquentes que les données normales, y compris dans les annotations. Ce déséquilibre n'est aussi généralement pas le même dans l'ensemble d'apprentissage que dans la tâche finale en test. Ce cas particulier de classification supervisée sera abordé dans les chapitres IV et sec. V.

détection non-supervisée

Le second cas de détection d'anomalies est celui où ces dernières sont totalement inconnues, trop rares ou trop variées pour pouvoir les annoter. Il n'y a alors pas d'annotations et la tâche de détection d'anomalies devient non-supervisée. Dans ce cas, la tâche s'apparente plutôt à de la détection de nouveauté ou de données aberrantes (*outlier*). De ce fait, la notion d'anomalie est à comprendre plus comme un concept d'anormalité, la normalité étant définie par les données de l'ensemble d'apprentissage. Bien que l'évaluation de telles méthodes est difficile (la notion d'anomalie n'étant pas définie), ces méthodes sont utiles pour déclencher des alarmes qui seront remontées à des opérateurs humains. C'est exactement le cas d'application du projet VANAHEIM qui finançait mon postdoc et dans lequel ces méthodes étaient utilisées pour choisir quelles prises de vues montrer à un opérateur humain qui surveillait en permanence un réseau de caméras dans les stations de métro de Turin.

les 3 sigmas

Une méthode générale, à la fois simpliste et fondamentale pour la détection d'anomalies, mérite ici d'être abordée en détails. C'est la méthode dite des 3σ (trois sigmas). Il s'agit de considérer le cas très simple d'une donnée scalaire, par exemple une mesure de latence réseau, une mesure de quantité de vibration sur un objet tournant ou une mesure de température en un lieu. La méthode des 3σ est très simple pour détecter les valeurs anormales : on mesure la moyenne et l'écart type empiriques de nos données « normales ». Puis, pour détecter les anomalies, on regarde si une nouvelle mesure s'éloigne de la moyenne de plus de trois fois l'écart type.

modélisation, estimation, vraisemblance

Une analyse plus précise de cette méthode des 3σ fait apparaître 3 étapes fondamentales qui sont sous-jacentes à de nombreuses approches de détection d'anomalies :

1. *supposer un modèle de génération des données* : dans le cas simple, on a supposé que nos mesures proviennent d'une loi normale de moyenne et écart type inconnus

2. *estimer/apprendre les paramètres du modèle génératif* : dans le cas simple, on a pris la moyenne et variance empirique de notre ensemble de données,
3. *juger de la vraisemblance des nouvelles données par rapport au modèle* : dans le cas simple, on a décidé que (statistiquement et si le modèle est correct) 0.2% des données seraient considérées comme anormales (pour une loi normale, il y a environ 0.1% des données qui s'éloignent de plus de 3 écarts types de chaque côté de la moyenne).

3σ avec des modèles probabilistes

Sous l'angle de cette méthode générique en 3 étapes, les modèles probabilistes sont une famille d'approches tout a fait adaptée pour la détection d'anomalies. En effet, les modèles probabilistes servent naturellement à l'étape de spécification du processus de génération, toutes les méthodes d'estimation et d'inférence probabiliste servent à l'estimation (au sens large) des paramètres mais aussi au calcul d'une vraisemblance pour de nouvelles données. Vu autrement, un modèle probabiliste permet de spécifier une structure pour l'expression d'une densité de probabilité, contrôlée par des paramètres. Cette densité de probabilité, combinée à un jeu de données, donne une fonction à optimiser pour l'estimation de paramètres. Cette même densité de probabilité, étant données des valeurs de paramètres (de manière déterministe ou stochastique), permet d'évaluer la vraisemblance de nouvelles données.

III.2 Décomposition non-supervisée en motifs récurrents

abstraction des données temporelles

Dans un but de détection d'anomalies mais aussi de fouille et d'analyse de données, nous avons développé de nouveaux modèles probabilistes pour les données temporelles assimilables à celles issues de vidéos. Pour ce faire, nous avons généralisé le problème en se découplant de la partie traitement d'image. Un second but était de pouvoir réutiliser les terminologies et de pouvoir construire sur les approches classiques liées aux topic models telles que *latent Dirichlet allocation* (LDA).

problème = fouille de motifs récurrents

Le problème peut être formalisé de la façon suivante : *étant donné un ensemble de documents temporels, trouver les motifs récurrents et leurs occurrences dans chacun des documents.* Autour de cette formulation, nous avons différents concepts clés :

- Un **document temporel** est un tableau à deux dimensions, avec un axe temporel et un axe de features. En référence aux topic models, les features sont appelées **mots**, l'axe correspondant ainsi aux mots d'un vocabulaire prédéfini. Une case *(mot, temps)* du document temporel contient le nombre de fois que le mot apparaît à cet instant de temps.

- Le **vocabulaire** définit la représentation des documents temporels. Il pourrait pour des documents textuels être un vrai vocabulaire de mots. Une variété de vocabulaires permet d'appliquer ces méthodes sur différents types de données. Les topic models classiques ont par exemple été appliqués dans un contexte de représentation d'image, grâce à des vocabulaires dits visuels, obtenus par exemple à partir d'un clustering de descripteurs tels que SIFT.
- Un **jeu de donnée** est un ensemble de documents temporels (potentiellement de longueurs différentes).
- Un **motif (temporel)** peut être vu comme une sous partie d'un document temporel. C'est un tableau 2D avec un même axe de features et une taille temporelle limitée, par exemple 20 pas de temps. L'axe temporel dans un motif est interprété comme un temps relatif par rapport à l'instant d'occurrence de ce motif.
- Une **occurrence de motif** définit quand apparaît celui ci : c'est une référence vers un motif, accompagnée d'un instant de temps à l'intérieur d'un document temporel. Une occurrence peut aussi contenir une information d'intensité (ou poids) de présence du motif.

La figure III.1 illustre la formulation et les concepts clés.

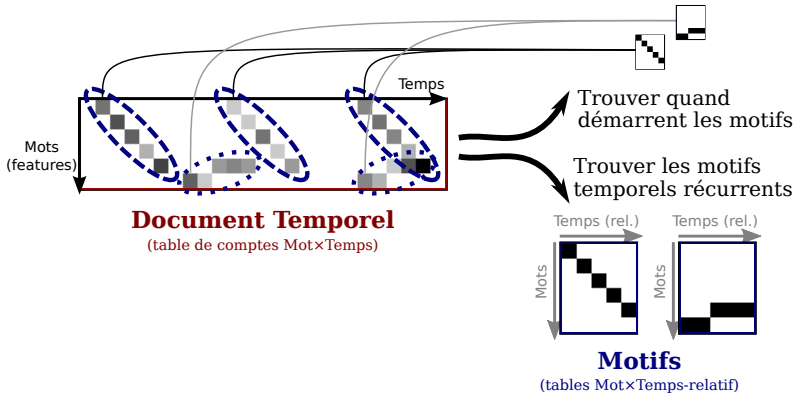


Figure III.1: Illustration d'un unique document temporel contenant 3 occurrences d'un motif (barre diagonale) et deux occurrences d'un autre motif (petite marche). Dans cet exemple, le vocabulaire est composé de 6 mots et les motifs durent 5 pas de temps.

explication avec la décomposition de vidéo

Ce problème de décomposition automatique de documents temporels en un ensemble de motifs récurrents et leurs occurrences permet non seulement la détection d'anomalies mais plus généralement l'analyse de données temporelles. Dans le cas de vidéos de carrefour routier, on peut extraire le mouvement (au niveau de chaque pixel et pour chaque image) pour former un vocabulaire, et la problématique revient alors à trouver automatiquement, à partir de quelques heures de vidéos, les groupes de pixels qui bougent ensemble et de manière cohérente dans le temps. On cherche à retrouver

les activités (ou motifs) inconnues et sous-jacentes, qui ont causé nos observations (mouvement). Le passage d'une voiture crée par exemple un motif spatio-temporel de pixels en mouvement qui est très caractéristique en fonction de sa trajectoire.

données non-négatives et additivité

Une supposition et une spécificité de cette formulation sous forme de mot et de mélange d'occurrences de motifs est que les observations sont des quantités de mot (donc positives) et que l'on suppose une forme d'additivité des données, dans le sens ou deux occurrences de motifs qui se superposent vont (approximativement) s'ajouter dans les observations.

III.3 Prise en compte du temps dans les topic models

Cette section est relativement longue car elle pose les bases communes pour la suite.

topic models (modèles à thèmes)

Le développement des contributions exposées dans ce chapitre a débuté à mon arrivée en postdoc à l'Idiap. Je me suis retrouvé à collaborer avec Jagannadan Varadarajan (doctorant de Jean-Marc Odobez), puis informellement à participer à son co-encadrement au quotidien. La base de ces contributions est d'étendre les topic models à des documents temporels. Il existe deux modèles de base dans la famille des topic models. Le premier est PLSA (*probabilistic latent semantic analysis*) qui suppose l'existence de thèmes sous-jacents (*latent topics*), inconnus, et que chaque document textuel (sac de mots) est un mélange de certains de ces thèmes. Le second est LDA (*latent Dirichlet allocation*), qui est une version bayésienne de PLSA, incluant ainsi la notion d'a priori sur l'ensemble des variables aléatoires considérées.

extension des topic models

La première contribution est le **modèle PLSM** (*probabilistic latent sequential motifs*, [34], [47]). C'est une extension de PLSA aux documents temporels. Les efforts précédents pour utiliser les topic models sur des données temporelles étaient principalement de deux natures :

HMM sur LDA

- Ajouter un modèle de type HMM (*hidden markov model*) au dessus des distributions de topics, pour modéliser le fait que les thèmes utilisés varient régulièrement dans le temps. Ceci est très pertinent pour la modélisation d'un corpus de bibliographie scientifique par exemple, mais un peu moins pour capturer une structure temporelle fine et locale. C'est typiquement ce qui est fait dans l'approche TOS-LDA (*temporal order sensitive LDA*) [55].

LDA sur fenêtre glissante

- Utiliser une fenêtre glissante sur un document temporel pour avoir un ensemble de documents (chacun étant une tranche d'un document temporel) et y appliquer PLSA ou LDA. Ceci est très adapté à l'objectif et

permet de dé-mélanger plusieurs occurrences simultanées. Par contre, comme chaque vrai motif (activité) est coupé à différents instants par la fenêtre glissante, le modèle doit multiplier les topics pour que chacune des nombreuses coupures possibles soit couverte. L'interprétation du résultat nécessite aussi de prendre ces redondances en compte, ce qui n'est pas forcément évident.

motifs avec instants d'apparition

Notre modèle PLSM adopte une formulation convolutionnelle et introduit le concept d'occurrence, bien qu'il ne soit alors pas formulé exactement en ces termes. Les paramètres à apprendre sont les motifs, de la même forme que, par exemple, par TOS-LDA, et les instants où ces motifs commencent, sous forme d'une distribution de probabilité sur l'espace produit motif×instant.

formulation de LDA/PLSA

Pour donner une formalisation rapide de la modélisation faite par PLSM, nous rappelons celle faite par LDA (et PLSA). En tant que modèle probabiliste, la modélisation consiste à donner une formule de vraisemblance, dans laquelle nous insisterons sur ce que sont les paramètres (ce qui est appris). LDA modélise un ensemble de sacs de mots, formés par une fonction (table) de compte d'apparition dans chaque document d, de chaque mot w, noté $n_{count}(d, w)$. LDA (de même que nos approches) suppose généralement le nombre de documents connu, de même que le nombre de mots dans les documents. Ceci n'a aucune influence, mis à part dans le cas où nous voudrions générer de nouveaux documents. La formulation de PLSA définie (exprimée en log-vraisemblance) par :

$$\log p(n_{count}) = \sum_{d=1}^{\#d} \sum_{w=1}^{\#w} n_{count}(d, w) \ \log \sum_{z=1}^{\#z} p_1(z|d) p_2(w|z).$$

Avec quelques éléments de notation et points clés suivants :

- pour éviter de multiplier les notations/variables dans ce document, $\#d$ le nombre de documents, $\#w$ la taille du vocabulaire, $\#z$ le nombre (fixé) de topics,
- $p_1(z|d)$ la probabilité du topic d'indice z conditionné au fait que l'on est dans le document d… il y a ici beaucoup d'implicite, il faudrait plutôt écrire $p_1(topic = z|document = d)$ ou alors $p_1(z = k|d = i)$ (en utilisant z et d comme variables aléatoires et k et i pour parcourir leurs domaines) mais ce raccourci est très présent dans la communauté des modèles probabilistes,
- p_1 et p_2 sont les (ensembles de) paramètres à apprendre, et contrairement à la convention de la communauté qui nous amènerait à les noter tous les deux p (sans indice), le choix est fait ici de les distinguer pour faciliter la lecture.

- les deux premières sommes sont pour couvrir l'ensemble des observations (log d'un produit de termes indépendants), la troisième est elle une somme sur la variable latente « topic », qui fait la complexité de ce modèle de mélange. Il faut déterminer pour chaque document les proportions des « topics » qu'il contient.

formulation de PLSM

En travaillant avec des documents temporels $n_{count}(d,t,w)$ et en introduisant la notion de temps tr relatif au début d'un motif et de temps ts de début d'une occurrence, une des formulations de PLSM fait apparaître une quintuple somme (trois sommes pour les observations, deux sommes pour les variables latentes) :

$$\log p(n_{count}) = \sum_{d=1}^{\#d} \sum_{t=1}^{\#t} \sum_{w=1}^{\#w} n_{count}(d,t,w) \log \sum_{z=1}^{\#z} \sum_{ts=1}^{\#ts} p_3(z,ts|d) p_4(w, tr = t - ts|z)$$

trouver les motifs et occurrences

On y vise à apprendre/inférer p_3, les distributions d'occurrence des motifs (une distribution pour chaque document temporel), et p_4, les motifs eux-mêmes, c'est-à-dire la distribution des mots dans le temps relatif (une distribution pour chaque motif). Ici aussi, les notations se détachent de celles utilisées dans les articles pour aider à la compréhension : principalement, les articles utilisent p (à la fois pour p_3 et p_4), mais aussi PLSM décompose par exemple $p_3(z,ts|d)$ en $p_{3a}(z|d)p_{3b}(ts|z,d)$ mais la formulation proposée ici est plus simple et surtout plus uniforme pour aborder les autres contributions.

EM pour PLSM

Ce modèle est un modèle de mélange. Un peu à la manière des GMM (*gaussian mixture models*), nous avons proposé dans [34] un algorithme de type EM (*Expectation Maximization*) qui permet d'optimiser le maximum de vraisemblance mais aussi le maximum a posteriori dans le cas où le modèle PLSM est augmenté d'un a priori.

L'étape E de l'algorithme EM va, pour un GMM, calculer/maintenir une estimation de la distribution des variables latentes, c'est-à-dire pour chaque point, son degré d'appartenance à chaque composante du mélange (à partir de leurs positions courantes), que l'on pourrait noter $p_{EM-GMM}(z_i|x_i,\varphi)$ (où φ sont les paramètres « globaux » des composantes du mélange). Pour faire l'équivalent avec PLSM, on peut regrouper les observations identiques et s'intéresser donc à chaque mot du vocabulaire w, temps d'apparition t à l'intérieur de chaque document d. Pour chacune de ces observations, il faut alors calculer son degré d'appartenance à tous les instants d'occurrence possibles (ts) de tous les motifs (z), que l'on pourrait noter comme $p_{EM-PLSM}(ts,z|d,t,w,\varphi)$.

L'étape M de l'algorithme EM, pour un GMM, utilise les degrés d'appartenance de chaque point à chaque composante pour ré-estimer les

paramètres φ (centre, covariance et poids des composantes) à partir de moyennes/covariances/comptes pondérés. De la même façon, EM pour PLSM va ré-estimer les paramètres globaux que sont les motifs et leurs instants d'occurrence, c'est-à-dire les distributions p_3 et p_4.

initialisations multiples

L'algorithme EM trouve un optimum local et il est donc nécessaire de le lancer plusieurs fois avec différentes initialisations pour trouver un optimal (plus) global. C'est le cas, tant pour un GMM que pour PLSM. La vraisemblance donne directement le critère de choix entre les paramètres obtenus avec différentes initialisations.

III.4 Contrainte de parcimonie dans les topic models

a priori parcimonieux

Cette contribution a en fait été créée et utilisée dans PLSM [47]. Sous forme de régularisation (ou a priori), elle permet de rendre parcimonieux les instants de départs (les distributions $p_3(z, ts|d)$). Dans PLSM, cette régularisation permet par exemple de limiter le nombre d'occurrences de motifs. En effet, la marche de la figure III.1 pourrait être expliquée non pas par une occurrence du motif « marche » mais par 3 occurrences successives d'un motif « dernier mot » suivi de 2 occurrences successives d'un motif « avant dernier mot ». Dès lors qu'il y a une variété de motifs par rapport à la taille du vocabulaire, ce cas de motifs contenant très peu d'information temporelle est en fait une solution optimale pour le modèle PLSM (sans régularisation) et survient en pratique. La régularisation permet alors d'obtenir des résultats utiles et capturant vraiment des informations temporelles.

utile pour tout type de distribution

Cette régularisation peut être appliquée pour différentes distributions, pas forcément temporelles. Dans le but d'insister sur cette contribution, mais aussi de partager cette formulation avec des communautés non intéressées par les aspects temporels, nous avons explicitement exposé cette contribution dans un workshop [35] et en tant que chapitre de livre [51].

éviter d'être uniforme

L'objectif est de rendre parcimonieuse une distribution (discrète) par l'ajout d'un terme de régularisation. Une formulation serait de minimiser l'entropie de la distribution mais cette approche peut s'avérer difficile à intégrer dans l'optimisation via EM. La formulation que nous avons choisie est de maximiser la divergence (au sens de la divergence de Kullback-Leibler) entre la distribution uniforme U et la distribution d'intérêt, plus précisément de minimiser, pour chaque document d : [ ]

$$KL\left(U \mid\mid p_3(z, ts|d)\right) = const + \sum_{z=1}^{\#z} \sum_{ts=1}^{\#ts} \frac{1}{\#ts\#z} \log(p_3(z, ts|d)).$$

formulation inadaptée ?

Deux spécificités auraient pu, à première vue, mettre en échec cette formulation :

- Le fait que la divergence de Kullback-Leibler soit plutôt adaptée pour des distributions qui sont relativement proches (la divergence de Kullback-Leibler pouvant vite devenir infinie) alors qu'ici l'intention est que la distribution soit très loin de l'uniforme. L'utilisation par ailleurs d'a priori, qui évitent les cas tels que des probabilités exactement égales à 0, est la raison qui fait que nous évitons les cas pathologiques où la divergence de Kullback-Leibler perd son sens.
- Le sens de la divergence de Kullback-Leibler utilisée où U et p_3 sont inversés par rapport au sens « habituel » présent par exemple dans les a priori en inférence variationnelle. Le choix est d'une part guidé par le fait qu'utiliser l'autre sens referait apparaître un terme d'entropie plus difficile à optimiser. En conséquence, on cherche donc à maximiser l'erreur d'approximation de la distribution uniforme pour la distribution p_3, ce qui en soit a bien du sens : pour l'ensemble (uniforme) des possibles, on veut qu'il soit difficile à coder par p_3 (donc que la valeur de p_3 soit la plus petite possible), le log faisant que l'on préfère une grande valeurs que plusieurs valeurs moyennes (toutes les valeurs ne peuvent pas être petites car p_3 est une distribution et doit donc sommer à 1).

mais pragmatique

Bien qu'imparfaite par plusieurs aspects, cette formulation est très facile à intégrer dans l'optimisation et s'est montrée efficace pour rendre les distributions parcimonieuses.

III.5 Modèles temporels bayésiens non-paramétriques

choix du nombre de composantes, modèles non-paramétriques

Une des problématiques courantes dans les modèles de clustering, de mélange comme les GMM ou les topic models est le choix du nombres de clusters ou composantes. Ce nombre est habituellement un hyper-paramètre pour lequel différentes valeurs sont testées. Ceci n'est pas totalement satisfaisant car cela nécessite des critères de sélection de modèles permettant de choisir a posteriori entre des modèles de dimensions différentes. Une famille d'approches pour intégrer ce choix dans la formulation du problème et son optimisation est celle des méthodes dites non-paramétriques. Nous nous intéresserons ici en particulier aux modèles bayésiens non-paramétriques.

HDLSM, un PLSM non-paramétrique

La premières contribution que nous avons proposée dans ce sens est le modèle HDLSM (*Hierarchical Dirichlet Latent Sequential Motifs*) [36] aussi nommé TAMM (*Temporal Analysis of Motif Mixtures*) quand nous le faisons évoluer vers VLTAMM (*Variable Length TAMM*) [48]. Par l'adaptation des modèles de type HDP (*Hierarchical Dirichlet Processes*, expliqués ci-dessous) aux données temporelles, HDLSM permet de ne pas spécifier directement le nombre de motifs, mais aussi de traiter naturellement les occurrences de motifs (à l'intérieur d'un document) de manière parcimonieuse, sans nécessité de régularisation telle que présentée dans la contribution précédente.

vers des GMM infinis et au delà

Dans ce qui suit, je souhaite prendre le temps d'expliquer, « avec les mains », le fonctionnement d'un modèle de mélange infini, que l'on nommera aussi DP (*Dirichlet Process*, ou processus de Dirichlet), qui est à la base des modèles bayésiens non-paramétriques tels que HDP ou HDLSM. Cette explication se base sur le modèle DP-GMM, donc un mélange infini où les composantes sont des gaussiennes. Pour aller plus loin (avec plus de formalisation) sur les aspects hiérarchiques et non paramétriques, le lecteur peut se référer à notre article [48] [ ] .

III.5.1 Vers des mélanges infinis

approche bayésienne pour les GMM

Dans un modèle de mélange de gaussiennes classique, il est nécessaire de choisir le nombre de gaussiennes, que l'on nommera K. Les paramètres à estimer sont alors les K moyennes des gaussiennes, leurs K matrices de covariance et leurs K poids habituellement notés $\{\pi_k\}_{k=1}^{K}$, sous la contrainte que π somme à 1. Ce sont ces paramètres qu'un algorithme tel que EM cherche à trouver. Si l'on imagine une situation avec peu d'observations ou avec un K grand, il peut y avoir un trop faible nombre de points pour estimer robustement les poids, mais surtout les moyennes et encore plus les covariances des gaussiennes. C'est pourquoi les approches bayésiennes posent des a priori sur ces paramètres (qui d'un point de vue optimisation peuvent souvent se ramener à des régularisations) pour formaliser le fait qu'un paramètre estimé avec peu d'information reste partiellement incertain. Par exemple, 5 points dans un espace à 8 dimensions ne sont pas suffisants pour estimer une covariance, mais ces 5 points donnent déjà une idée de la tendance que pourrait avoir la vraie covariance à dévier d'une covariance identité.

même a priori pour les moyennes et variances

L'idée d'un DP-GMM est relativement simple et consiste à pousser le nombre de composantes K à l'infini. Il devient alors indispensable de traiter le problème de manière bayésienne puisque le nombre de paramètres est (infini-

ment) plus grand que le nombre d'observations. En ce qui concerne chaque moyenne et matrice de covariance de l'infinité de gaussiennes, leurs a priori peuvent être définis exactement de la même manière que pour un mélange bayésien fini.

a priori sur une infinité de poids

Pour le vecteur de poids π, qui est maintenant infini, il n'est plus possible d'appliquer ce qui est fait dans le cas fini (généralement une distribution uniforme) car il n'y a pas de sens à définir une distribution uniforme sur un espace (discret) infini. À la manière du « paradoxe » d'Achille et la tortue (séries convergentes), il est pourtant possible de définir un a priori (un processus stochastique) régissant le vecteur infini de poids π, tel qu'il y aie une infinité de poids non nuls mais sommant à 1. Ce principe est formalisé par le processus appelé SBP (*stick breaking process*) ou GEM (du nom des auteurs *Griffiths, Engen, McCloskey*), qui est contrôlé par un seul hyper-paramètre β (il existe aussi les processus de *Pitman-Yor* qui ont deux paramètres, pour contrôler la queue de la distribution π obtenues). Le principe du SBP est de commencer avec un bâton de longueur 1, puis de répéter pour $k = 1$ à ∞ les étapes suivantes :

- tirer un nombre ν_k aléatoirement dans l'intervalle $[0, 1]$ (plus précisément selon une loi Béta de paramètres 1,β)
- casser le bâton en deux à une proportion ν_k
- la longueur du premier morceau définie le poids π_k (qui vaut donc ν_k fois la longueur du bâton, donc $\nu_k \cdot \prod_{i=1}^{k-1}(1 - \nu_i)$
- le morceau restant est conservé pour la suite du processus, sa longueur étant $(1 - \nu_k)$ fois la longueur du bâton, donc $\prod_{i=1}^{k}(1 - \nu_i)$).

impact de la concentration β

Intuitivement, si β est petit, il y a une grande probabilité de tirer un ν_k grand (de part la distribution Béta). On va donc avoir tendance à mettre de coté des grands morceaux du bâton, c'est à dire à créer rapidement des composantes avec un poids fort. Le paramètre β est souvent appelé *concentration* car il contrôle la « densité » de composantes : β petit donne peu de composantes avec des poids forts, β grand donne beaucoup de composantes de poids proches [ ] . Ce processus ne définit qu'un a priori (et non pas une contrainte forte sur les valeurs) qui entre dans un compromis avec l'information venant des données.

formulations équivalentes d'un DP

Le processus de Dirichlet (DP) n'est rien d'autre qu'un a priori sur un modèle de mélange infini dont les poids sont tirés d'un SBP. Autrement dit et pour résumer les différentes formulations : pour un mélange infini G de composantes (par exemple gaussiennes) notées $Comp(\varphi_k)$, dont les paramètres (de chaque composante) φ_k suivent un a priori G_0, si on considére que G suit un a priori de type processus de Dirichlet, on peut écrire de manière

équivalente :

$$
\begin{array}{l|l|l}
G \sim DP(\beta, G_0) & \pi \sim GEM(\beta) & \forall k \\
 & \forall k, & \quad \nu_k \sim Beta(1, \beta) \\
 & \quad \varphi_k \sim G_0() & \quad \pi_k = \nu_k \cdot \prod_{i=1}^{k-1}(1-\nu_i) \\
 & G = \sum_{k=1}^{\infty} \pi_k Comp(\varphi_k) & \quad \varphi_k \sim G_0() \\
 & & G = \sum_{k=1}^{\infty} \pi_k Comp(\varphi_k)
\end{array}
$$

(ou $\sim$ peut se lire « suit la loi » ou « est tiré de », selon les points de vue)

III.5.2 Optimisation de mélanges infinis

MCMC ou approches variationnelles

L'optimisation la plus simple pour ces types de modèles de mélange infinis est l'échantillonnage de Gibbs, une méthode de la famille MCMC (*Markov Chain Monte-Carlo*). Des approches variationnelles capables de gérer cet infini ont cependant été développées et peuvent s'appliquer ou s'adapter à toutes ces familles de modèles.

échantillonage de gibbs

Pour un modèle de mélange classique, le cœur de l'échantillonnage de Gibbs consiste d'abord à s'initialiser en affectant chaque observation à une composante aléatoirement. Ensuite un processus itératif considère chaque observation en boucle et reconsidère son affectation étant données les affectations de toutes les autres observations fixées (à leur valeur courante dans l'algorithme). Bien que généralement coûteux, l'échantillonnage de Gibbs a des avantages théorique (convergence au sens des méthodes MCMC) et peut s'avérer plus efficace dans le cas de données parcimonieuses (par exemple dans notre cas, pour des documents temporels constitués essentiellement de zéros).

MCMC et mélange infini

Le principe pour gérer un mélange infini et de se rendre compte qu'à un instant donné dans l'optimisation, étant donné un nombre d'observation N fini, nous n'avons de l'information que sur un nombre fini de composantes (au maximum N et généralement bien moins). Ainsi, l'échantillonnage peut (à chaque fois qu'il reconsidère une affectation) gérer d'un côté les composantes existantes (c'est-à-dire auxquelles sont actuellement affectées des points), de manière classique. L'ensemble (infini) de toutes les autres composantes est lui géré de manière analytique. En effet, toutes ces composantes peuvent être regroupées car elles sont identiques dans le sens où aucun point n'apporte d'information sur ces composantes (seul l'a priori informe de la forme possible de ces composantes). La réaffectation dans le processus de Gibbs se fait donc en mettant en compétition les composantes existantes et le cas d'une « nouvelle composante ».

III.5.3 Modèles hiérarchiques (de mélanges infinis)

HDP, mélange infini de groupes

L'évolution hiérarchique du modèle infini, HDP-GMM (*Hierarchical Dirichlet Processes for GMM*), modélise un autre type de données qui sont des groupes d'observations. Pour un mélange simple, type DP-GMM, prenons l'exemple où chaque point est un son d'une seconde enregistré à un endroit dans la nature. À la manière d'un clustering, un modèle de mélange va essayer de trouver les groupes de sons qui se ressemblent. Le cas hiérarchique HDP-GMM s'intéresse au cas où nous aurions plusieurs lieux d'enregistrements de sons. Le but est alors de trouver les groupes de sons qui se ressemblent, mais en intégrant le fait que les groupes présents et leurs importances relatives varient d'un site d'enregistrement à l'autre : chaque site à ses propres proportions (possiblement nulle) de félins, d'oiseaux, d'insectes, etc. Dans les topic models, c'est exactement le même principe : il y a des thèmes globaux et chaque document est composé d'une sous-partie (pondérée) de ces thèmes. La seule différence est que les topic models sont des mélanges de lois discrètes (loi multinomiales) au lieu de gaussiennes (loi normales). L'évolution bayésienne de PLSA est LDA (latent Dirichlet allocation) et son évolution non-paramétrique est souvent appelée simplement HDP.

III.5.4 Le modèle HDLSM (ou TAMM) : infinité de motifs

HDLSM = HDP avec translation temporelle

Nous avons (enfin) tous les éléments pour introduire le cœur du modèle HDLSM [36], la version non-paramétrique de PLSM. Comme attendu, HDLSM suppose un ensemble de motifs qui sont, comme dans PLSM, des tables de probabilités $p(w, tr|z)$ mais en nombre infini. La différence plus originale est que, pour chaque document temporel, ce n'est pas seulement une sous-partie de ces motifs qui est tirée (comme le ferait un HDP ou HDP-GMM). Un document est plutôt un ensemble d'occurrences de motifs (donc indirectement une sous-partie des motifs) qui contiennent aussi une information de position de démarrage du motif dans le document. Ce n'est qu'après la publication [36] que nous avons découvert en refaisant de la bibliographie avec un œil nouveau que ce concept est assez similaire aux *Transformed DP* [56].

infinité d'occurrences d'une infinité de motifs

Formulé autrement, HDLSM modélise « simplement » chaque document comme un mélange infini (et donc parcimonieux par construction) d'occurrences de motifs (avec un instant d'occurrence et un indice de motifs) qui font référence à un ensemble infini, global, de motifs. L'avantage sur PLSM est donc de trouver automatiquement un nombre de motifs pertinents pour expliquer les données observées, mais aussi de gérer les occurrences de manière plus sémantique (avec une parcimonie présente par construction). HDLSM a donc deux hyper-paramètres de concentration

pour ses DP : γ qui influence le nombre global de motifs et α qui influence la densité d'occurrence dans les documents temporels.

a priori et échantillonage par bloc

L'implémentation que nous avons proposée optimise HDLSM à l'aide d'un échantillonnage de Gibbs (méthode MCMC). Pour aider le modèle à bien apprendre (ou plus correctement pour accélérer la convergence en pratique de la méthode MCMC), nous avons mis en place une étape de ré-echantillonage des variables en bloc (*blocked Gibbs sampling*) qui en plus travaille sur un sous espace. Ceci pourrait s'apparenter à un changement de variable. L'idée est de ré-échantillonner simultanément tous les instants de départs des occurrences d'un même motif. Cela est utile car l'initialisation peut faire qu'un motifs du modèle « coupe » un vrai motif (une activité). Par exemple, le motif capture du vide puis le début de l'activité. En pratique, il est quasi impossible pour le Gibbs sampler de sortir de cette situation, bien qu'en théorie cela arrivera. C'est pourquoi nous avons mis en place un a priori, sous forme d'une rampe décroissante, qui favorise le fait que le début des motifs ne soit pas vide, et cette étape de ré-échantillonnage par bloc. Le problème de motif coupé peut aussi apparaître de l'autre coté (au début du motif appris) et c'est pourquoi l'a priori donne un traitement particulier au premier pas de temps à l'intérieur du motif [ ] .

III.5.5 Le modèle VLTAMM : motifs à longueur variable

trouver la longueur des motifs

Une limitation de PLSM et HDLSM est que la durée (maximale) d'un motif doit être donnée en hyper-paramètre. Même si l'a priori présenté ci-dessus avec HDLSM permet d'une certaine façon de donner une longueur maximale assez longue et de récupérer les vrais motifs au début de long motifs, avec beaucoup de vide à la fin, cette situation n'est pas parfaite. Dans l'article qui présente VLTAMM [48], pour *Variable Length Temporal Analysis of Motif Mixtures*, en plus de consolider les différentes explications et expériences autour de HDLSM (appelé TAMM), nous avons aussi proposé un nouveau modèle qui gère des motifs de durées variables et déterminées automatiquement.

exponentielle tronquées en poids, distribution à support variable

Il nous a alors fallu proposer un a priori (sur les motifs eux mêmes) qui permet de gérer des distributions à support variable (plus le motif est long plus son support est grand). Contrairement au cas du mélange infini qui est un problème « résolu » (par les DP), le cas d'une distribution n'avait pas de solution classique à notre connaissance. Notre choix a été de proposer de remplacer l'a priori de HDLSM (la rampe décroissante) par une distribution exponentielle tronquée. La masse à partir de laquelle l'exponentielle est tronquée reste un hyper-paramètre et il contrôle principalement la pente de la rampe. Le paramètre λ (paramètre dit d'intensité pour la loi exponentielle)

contrôle alors la taille du support de la distribution tronquée. Un tel paramètre d'intensité est estimé pour chaque motif (et noté λ_k). Sans rentrer dans les détails de l'échantillonnage, les paramètres λ_k sont soumis à un a priori de type Gamma et sont échantillonnés grâce à un mélange original d'échantillonage par rejet et d'a priori conjugué [ ] .

III.6 Applications et gestion des dépendances à long terme

applications dans le projet VANAHEIM

Dans le contexte du projet VANAHEIM, nous avons publié des travaux plus directement liés à l'application de traitement multi-caméras et du choix de caméra à afficher dans un réseau de caméras [49], [40] mais aussi dans une conférence d'éthologie humaine [41] (qui consiste à étudier le comportement de l'être humain en tant qu'espèce animale, sans biais d'interprétation) organisée par nos partenaires du projet VANAHEIM.

applicabilité à une variété de données réelles

Au-delà des données vidéos et des données de synthèses qui étaient au cœur de la problématique des articles principaux présentés précédemment, nous avons aussi appliqué ce modèle dans différents contextes, parfois en proposant des améliorations de ce modèle. Nous avons appliqué les modèles sur des données audio capturées à partir de deux microphones (placés côte à côte), et transformées en documents temporels où l'axe des mots est un axe d'azimut (discrétisé). En effet, la différence de temps d'arrivée d'un son aux deux microphones indique la direction d'où le son vient. Ces expériences ont été directement incluses par exemple dans [48].

dépendances long terme et irrégulières

Une évolution importante du modèle PLSM est celle que nous avons faite avec Jagannadan Varadarajan dans le contexte de sa thèse. L'idée est que PLSM permet bien de capturer les activités récurrentes (par exemple un piéton qui passe) mais est peu adapté pour capturer les dépendances entre activités, si ces dépendances ne sont pas « dures ». Ainsi, si un piéton s'avançant de gauche à droite vers une vitrine (un premier motif) est souvent suivi d'un piéton partant vers la droite depuis la vitrine (un second motif) mais avec un temps entre ces deux motifs qui est variable, PLSM (et autres) ne pourront pas capturer cette co-occurrence facilement et de manière structurée. Le modèle MERM [42] (*mixed event relationship model*) utilise PLSM pour obtenir une représentation parcimonieuse d'une scène (sous forme d'occurrences de motifs) et ajoute deux concepts au dessus de cette représentation :

- un modèle global de type HMM (chaîne de Markov cachée) pour capturer l'aspect cyclique d'un scène, typiquement des scènes de trafic, régies par des feux de circulation,
- une notion de règles probabilistes (apprises) qui permet de capturer le fait qu'une observation (donc une occurrence de motif PLSM) ap-

paraît souvent après une autre activité avec un délai qui est modélisé comme une distribution normale sur le temps (dont le délai moyen et la variance sont des paramètres estimés).

données d'hydrographie

Nous avons aussi collaboré avec des colloborateurs nous ayant fourni des données d'hydrographie (pour étudier et prédire les crus et inondations). Comme il s'est avéré que le volume de données était limité et que l'alignement des événements pouvait être fait manuellement, nous avons utilisé LDA avec un vocabulaire incluant un aspect temporel plutôt que nos modèles temporels qui cherchent les instants d'occurrence des événements. Les publications associées sont [43], [44].

analyse de corpus, détection d'anomalie, multi-caméra

Ces modèles ont été utilisés dans le projet VANAHEIM pour de l'analyse de flux vidéo, de la détection d'anomalies et de la sélection de caméras parmi plusieurs caméras (pour montrer les plus anormales à un opérateur en direct). Un cas d'étude rapporté dans [37] mais aussi intégré aux expériences de [48] est celui de l'analyse jointe de plusieurs caméras, par exemple dans une station de métro. Sans aucune calibration ni supervision, la problématique est d'étudier les co-occurrences de mouvements dans une caméra mais aussi entre les caméras (une personne qui passe va créer un mouvement sur une caméra puis sur une autre etc).

PLSM sur GPU

J'ai dirigé le stage de [45] pour accélérer les modèles par l'utilisation de GPU (moins courant et facile à l'époque, fait d'ailleurs en OpenCL) et par une ré-écriture gérant bien la parcimonie dans les données d'entrée. Ceci était en particulier capital pour se passer de la première phase de réduction de dimension opérée pour PLSM (en pratique un PLSA/LDA, donc non temporel, est lancé sur les données brutes pour faire une pré-réduction de dimension, par exemple pour passer d'un vocabulaire de taille 60000 à un vocabulaire de taille 100 en entrée de PLSM). D'autres travaux, finalisés après mon départ de l'Idiap, se sont concentrés sur une implantation CUDA de PLSM [52].

reconnaissance d'activité, modèles continus

Dans le cadre de son postdoc, Romain Tavenard a repris en main et évalué ces modèles dans la reconnaissance d'action dans des vidéos [46]. Nous avions aussi expérimenté des modèles similaires mais totalement continus, avec un espace continu au lieu d'un vocabulaire, et un temps continu. Bien que rien n'ait été publié, cette exploration était très intéressante mais a dû être abandonnée de part nos recrutements respectifs en tant que maîtres de conférences et nos sollicitations de part et d'autre.

essais sur spectrogrammes

Les topic models temporels ont aussi été ponctuellement testés sur des spectrogrammes d'enregistrements audio sous-marins à l'occasion du stage de Valentina Zantedeschi. La collaboration avec l'ENES n'ayant alors pas encore démarré (voir section VII.3), ces travaux n'ont pas lancé de dynamique et nous n'avons pas poursuivi sur cette piste.

III.7 Suite : auto-encodeurs et détection d'objets

avantages et limites des modèles probabilistes

Les modèles probabilistes tels que PLSM ou HDLM, ont de nombreux avantages : ils sont bien fondés de part leur modélisation cohérente, à partir de probabilités, de bout en bout. Ils offrent de plus la possibilité d'introduire des variables latentes en leur donnant une sémantique via la structure du modèle, ils permettent d'appliquer différentes méthodes d'inférence, etc. Cependant, ces modèles sont souvent coûteux lors de l'inférence et leur passage à l'échelle peut poser problème (même si certaines formes d'inférence, comme l'inférence variationnelle stochastique, appliquées à certains modèles comme LDA arrivent à très bien passer à l'échelle).

III.7.1 Auto-encoders convolutifs

feed-forward contre explain-away

Dans le but d'explorer à quel point une approche de type auto-encodeurs pouvait remplir le même rôle qu'un modèle de type PLSM, nous avons encadré le stage de Kevin Bascol sur ce sujet, qui a donné lieu à une publication [18]. Kevin a d'ailleurs continué avec Élisa Fromont et moi-même dans le cadre d'une thèse CIFRE en deep learning, travaillant sur deux thèmes traités dans ce manuscrit (chapitres IV, sec. V). La question principale était de savoir ce qu'un réseau de neurones, qui est principalement de type *feed-forward* (donc n'itère pas pour réaliser une estimation/prédiction comme le ferait par exemple EM) pouvait arriver à réaliser par rapport à l'inférence des modèles probabilistes qui inclut fortement le concept du *explaining away* (qui semble se traduire par « expliquer » en français, mais le *away* me semple capital).

explaining away

Le concept de *explaining away* est très présent dans l'inférence d'un modèle probabiliste et en particulier de modèles hiérarchiques. Pour simplifier l'explication, on supposera que l'on connaît les motifs du modèle PLSM (cela est d'ailleurs le cas quand on a appris des motifs et que l'on essaie de les retrouver dans des nouvelles données). Imaginons qu'une même observation puisse s'expliquer par deux occurrences de motifs o_1 et o_2, alors avant tout raisonnement, un doute persiste sur quelle occurrence a causé cette observation. Cependant, si d'autres observations renforcent l'idée que o_1 est bien présente (mais pas particulièrement o_2), alors, pour l'observation considérée, la présence de o_1 *explains away* l'existence de o_2. Autrement dit, le renforcement de la croyance en la présence de o_1 explique la présence de l'observation et donc réduit le besoin pour o_2 d'exister.

base pour un auto-encodeur PLSM

L'idée pour émuler PLSM avec un auto-encodeur est en fait d'utiliser un auto-encodeur convolutif avec une seul couche de convolution pour l'encoder (et une pour le décodeur). Ainsi les filtres de (dé)convolution correspondraient

aux motifs et la représentation latente au instant de démarrage. Du coté décodeur, cette vision est parfaite : le document reconstruit correspond bien à une superposition additive de motifs. Le problème vient de l'encodeur qui est trop limité pour réaliser juste en une passe *feed forward*, l'équivalent de ce que l'inférence probabiliste réalise, et en particulier de gérer les situations de *explains away* qui sont présentes dans les documents temporels même les plus simples. Un autre aspect que ne gère directement pas un auto-encodeur est la notion de parcimonie.

contraindre et régulariser

Les contributions de ces travaux consiste en l'ajout de différents concepts dans un auto-encodeur convolutif pour le rendre le plus proche possible, dans son comportement, de PLSM. Nous avons tout d'abord contraint les motifs et les activations à être positives. Ensuite, pour favoriser la parcimonie, nous avons mis une régularisation de type ℓ_1 sur les poids de convolution (les motifs), et sur la représentation latente (les occurrences). Pour permettre dans une certaine mesure de trouver automatiquement le nombre de motifs, nous avons aussi introduit une régularisation *group-lasso* sur les filtres de convolution, les poids d'un filtre constituant un groupe, ce qui amène à faire disparaître les motifs dont l'intensité est faible.

nouvelles couches de seuillage adaptatif

Tout ces ajouts ne résolvent pas la question du *explains away*. Pour ce problème, en s'inspirant du cerveau qui a des connexions inhibitrices entre des neurones proches (l'activité d'un neurone peut donc réduire celle d'un autre neurone, tout comme le *explaining away* de variables aléatoires) nous avons proposé une version douce de suppression de non maximum local, en soustrayant aux activations (représentation latente) une version lissée par un filtre gaussien de ces même activations. Cette méthode est efficace mais ne fonctionne que pour des occurrences qui sont relativement proches (et selon la formulation, uniquement des occurrences d'un même motif). C'est pourquoi, nous avons proposé en plus une version d'inhibition plus globale sous la forme d'une couche appelée AdaReLU (*adaptive rectified linear unit*. Le principe est de ne conserver que les activations les plus fortes, plus précisément en forçant à 0 toutes les activations (d'un groupe) qui sont inférieures à un pourcentage de l'activation la plus grande (par exemple 60%).

III.7.2 Classification de séries temporelles : interprétabilité et early classification

interprétabilité pour les séries temporelles

Sur un thème proche mais avec des approches différentes, nous nous sommes intéressés dans la thèse de Yichang Wang, que j'ai co-encadré avec des collègues de Rennes, à l'aspect « interprétabilité » des modèles de classification de séries temporelles. Ces modèles convolutifs sont beaucoup plus performants (à la fois en précision et en tant de calcul) que les modèles dits à

base de *shapelets*. Cependant, les modèles à base de *shapelets* utilisent des morceaux de séries temporelles issus du jeu de données comme motifs et sont donc (considérés) interprétables alors que les réseaux convolutifs apprennent des filtres de convolution qui sont généralement (moins) interprétables. Bien que l'apprentissage était, dans ce contexte, supervisé (contrairement à la fouille de motifs), certains concepts se retrouvent entre les deux problématiques : données temporelles, utilisation d'approches convolutives, importance de la structure temporelle des motifs, besoin d'une sémantique claire sur ce que le modèle apprend.

régularisation adversaire, rendre les filtres proches des séries

Après plusieurs explorations, le cœur de la thèse de Yichang Wang a été de proposer une régularisation adversaire, que l'on peut expliquer en s'inspirant des GAN (*generative adversarial network*). Là où un GAN essaie de faire que les images que son générateur produit soient proches des images réelles, nous avons formulé le fait que nous voulions que les filtres appris soient proches de certains morceaux de séries du jeu de données. Nous avons donc proposé une architecture avec un réseau critique (basé sur une approche de type WGAN-GP (*wasserstein GAN with gradient penalty*) qui rapproche les filtres de morceaux de séries tirés au hasard. Notre modèle n'a pas de générateur puisque nous voulons contraindre les filtres *appris* (et non des choses *générées*). Nous avons aussi réalisé un tirage biasé des morceaux de séries utilisés par le réseau critique, de façon à rapprocher les filtres non pas de l'ensemble des sous-morceaux mais plutôt des sous-morceaux leur ressemblant le plus. Ces travaux ont été mis à disposition (et cités) d'abord sur arxiv [24] puis présentés à la communauté français [23] et finalement en conférence [25].

approche générique pour l'*early classification*

Lié aussi à la classification de séries temporelle, j'ai collaboré avec Marc Rußwurm et son équipe d'encadrants sur une problématique relativement nouvelle de early classification qui consiste à faire de la classification de séries temporelles (donc donner une étiquette à une série temporelle) mais en donnant la réponse le plus tôt possible, c'est à dire sans avoir vu l'ensemble de la série. Cette tâche s'applique à tous les cas où les données de la série arrivent progressivement et consiste nécessairement en un compromis entre réponse rapide et réponse juste : on veut généralement garder une excellente performance en classification tout en classifiant le plus tôt possible. Nous avons proposé une approche générale pour l'utilisation de modèles récurrents (LSTM par exemple) pour les apprendre à être bons en early classification. Le principe est que, à chaque pas de temps, le modèle prédit à la fois une confiance (probabilité qu'il faille prendre la décision) et une classe.

cité mais difficile à publier

Formaliser ces deux prédictions nous a permis de mettre au point une fonction de perte bien fondée qui permet d'apprendre le modèle de bout en bout, sans aucun pré-entraînement. Après un temps certain (et plusieurs re-soumissions) et plusieurs dizaines de citations sur arXiv [53], ces travaux ont finalement été acceptés dans une version orientée vers les applications de suivi temporel de cultures via des images satellites [54].

III.7.3 Auto-encodeurs variationnels profond pour l'image

décomposition non-supervisée d'images en objets

Via l'inférence variationnelle amortie, il y a un lien très fort entre les modèles probabilistes et les architectures de réseaux de neurones sous forme d'auto-encodeurs (pour plus de détails, voir une tentative d'explication accessible de ce lien entre modèles probabilistes et auto-encodeurs variationnels (VAE, *variationnal autoencoders*) que j'avais rédigée sous forme de billet sur un blog qui ne demande qu'à revivre sous une autre forme). Dans le contexte de l'image, des approches proches de PLSM ont été développées ces dernières années dont parfois dans une formulation sous forme de modèle probabiliste, parfois sous forme d'auto-encodeurs simples. L'idée revient souvent à avoir un encodeur qui réalise de la détection d'objets (beaucoup plus compliqué que dans les travaux de Kevin Bascol) et un décodeur qui fait l'équivalent d'un processus de synthèse d'image, tout cela étant appris de bout en bout sans étiquettes ou boîtes englobantes. À cela s'ajoute la problématique de segmentation objet/fond et la modélisation du fond. En collaboration avec Thierry Fournel de l'équipe image du laboratoire, je co-encadre Sayan Chaki dans le contexte du projet multidisciplinaire ANR ROIi sur l'analyse d'ornements anciens dans des livres imprimés. Une partie de sa thèse vise à développer des approches à base d'auto-encodeurs où l'espace latent encode une description fine des objets présent dans l'image (type, position, taille, etc).

Chapitre IV

Adaptation de domaine et apprentissage multi-tâches

Publications: [30] [29] [28] [27] [22] [5] [4] [57] [19] [3] [2] [1] [58].

Projets : TADALoT, MIVAO, Solstice.

Codes et liens divers

- Code pour les expériences de MLOT, publié à la conférence IJCAI 2020.
  https://github.com/twitwi/MLOT
- Code pour les expériences de OTT, publié à la conférence AAAI 2022.
  https://github.com/twitwi/Optimal_Tensor_Transport
- Exploration de transitions de slides par transport optimal faite par Tanguy Kerdoncuff pour sa soutenance.
  https://github.com/twitwi/TransitionPDF
- Vidéos du script de la démo pour l'HCÉRES sur le transport.
  https://home.heeere.com/data/videos/
- Exploration interactive de résultats pré-calculés avec l'outil de la démo.
  https://dl.heeere.com/perm/2021-OT-precomputed/

IV.1 Apprentissage non i.i.d., apprentissage par transfert

i.i.d.

En apprentissage automatique, il est très souvent fait la supposition que les données sont « i.i.d. », indépendantes et identiquement distribuées. Les données sont supposées provenir d'une distribution sous-jacente fixe, mais inconnue. Cette supposition permet d'utiliser des raisonnements statistiques pour étudier la qualité des algorithmes d'apprentissage, et en particulier pour prouver une convergence vers une bonne solution quand la taille du jeu de données augmente. Pour être plus précis, on parle ici de distribution sur l'espace $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ qui est l'espace produit entre l'espace des entrées $\mathcal{X}$ (valeur en entrée de la fonction à apprendre) et celui des étiquettes $\mathcal{Y}$ (valeur à prédire en sortie).

non-i.i.d

Cependant, il existe des situations où cette supposition n'est pas complètement valide. Parmi elles, on peut mentionner deux cas (non traités dans ce manuscrit). Le premier concerne les situations où les données (d'apprentissage) ne sont pas indépendantes, par exemple quand on observe des actions/transactions/achats dont plusieurs sont faites par un même utilisateur. Le second décrit un système qui, au moment du déploiement, est activement attaqué ou manipulé par des utilisateurs, par exemple par des fraudeurs (pour un algorithme de détection de fraudes) ou des manipulateurs d'opinions (pour un algorithme de recommandation ou de blocage de comptes frauduleux).

apprentissage par transfert

Les cas que nous considérons dans nos travaux sont ceux relevant de l'apprentissage par transfert. Cette catégorie regroupe l'ensemble des situations où les jeux de données d'entraînement et de test ne sont pas supposés issus d'une même distribution (généralement l'hypothèse i.i.d. reste valide dans chaque jeux de données). On suppose que l'on a un ensemble d'apprentissage, dit « source », et un ensemble de test, dit « cible ». Le but de l'apprentissage par transfert est d'arriver à réutiliser de la connaissance apprise sur le jeu source pour être (plus) performant sur le jeu cible.

adaptation de domaine

Un cas particulier d'apprentissage par transfert est l'adaptation de domaine. Dans ce scenario, ce qui change est la distribution sur $\mathcal{X}$ entre les jeux de données source et cible. C'est par exemple le cas si l'on apprend un détecteur de chats sur des images (sources) de dessins animés et que l'on veut l'utiliser sur des images (cibles) réelles. La distribution des images est clairement différentes entre le jeu source et le jeu cible. Par contre, la fonction reste globalement la même (ou compatible) : si on voit une image réelle (cible) qui ressemble à un dessin de chat (comme dans le jeu source), alors la décision doit être la même (c'est un chat). Le terme « adaptation de domaine »

vient du fait que l'on veut s'adapter à un changement dans les entrées de la fonction à apprendre, c'est à dire son *domaine*.

et multi-tâche

Bien que le titre de ce chapitre fasse référence à l'adaptation de domaine qui est un sous problème de transfert, il contient d'autres contributions en transfert. C'est le cas de l'apprentissage multi-tâches, où la spécificité est que la fonction à apprendre n'est pas la même en source et en cible. Par exemple, si l'on a un jeu de données (source) fait pour de la détection de véhicules, on pourrait vouloir ré-utiliser la connaissance que l'on peut extraire de ce jeu données pour réaliser une tâche (cible) de détection de voitures (mais pas les autres véhicules). Les tâches sont liées mais pas identiques, on peut donc s'attendre à ce qu'une tâche puisse aider à résoudre l'autre. Un scénario classique est celui où l'on a beaucoup de données sur la tâche source mais très peu sur la tâche cible. Une méthode très commune avec le deep learning est d'utiliser une sous-partie d'un réseau pré-appris, par exemple sur ImageNet, comme base pour réaliser une nouvelle tâche de traitement d'image. Cette approche de *fine-tuning* a l'avantage de ne pas utiliser les données sources pour le transfert, mais uniquement le modèle appris sur ces données dernières.

adaptation de domaine non-supervisée (UDA)

Une tâche d'adaptation de domaine courante est celle dite non-supervisée (*unsupervised domain adaptation*, UDA), décrite généralement par des données sources étiquetées et des données cibles non-étiquetées. Le but est alors d'être le plus performant possible sur la distribution des données cibles. On parle d'une situation d'apprentissage transductif si les données test (cibles) sont disponibles dès l'apprentissage du modèle. Le cas non-transductif est celui où l'on ne peut pas utiliser les données cibles pendant l'apprentissage (ou vu autrement, on n'a plus accès aux données sources au moment de l'adaptation). D'une manière générale, l'adaptation de domaine non-supervisée est un cadre compliqué, tant d'un point de vue théorique que d'un point de vue pratique.

problème : non-adaptabilité et inconnue λ

Un obstacle théorique de l'UDA est le fait que rien ne dit que l'adaptation soit possible, puisqu'il n'y a pas de supposition explicite particulière (de type données i.i.d.). Si les jeux de données (ou tâches) sont très différents alors il se peut que l'adaptation soit impossible et donne lieu à des performances pires que le hasard. C'est pourquoi de nombreux résultats théoriques dans ce cadre non-supervisé font apparaître un terme de non-adaptibilité (souvent noté λ). Ce terme λ est inconnu, impossible à estimer sans étiquettes cibles et souvent supposé petit pour que l'adaptation puisse avoir lieu. Intuitivement, λ peut mesurer l'erreur que ferait le meilleur classifieur sur l'ensemble des distributions sources et cibles, autrement dit, à quel point il est difficile de bien résoudre les deux problèmes.

problème : validation croisée non-supervisée

Un obstacle plus pratique de l'UDA est celui du choix d'hyper-paramètres par validation croisée. Comme il n'y a pas d'étiquettes cibles, il n'est pas possible de mettre de côté un ensemble de validation pour évaluer un jeu d'hyper-paramètres. Il faut donc une solution pour évaluer ces derniers sans étiquettes (même si certains protocoles expérimentaux de papiers publiés ont utilisé à tort les données de test pour cela). Une solution est de ne pas avoir d'hyper-paramètres ou de proposer des critères automatiques pour les fixer (sans utiliser les étiquettes).

validation croisée inversée

Une solution générique est d'utiliser une procédure de *validation croisée inversée*. Le principe est d'appliquer la méthode d'adaptation de domaine non-supervisée pour étiqueter les données cibles. En utilisant ces données cibles maintenant (pseudo-)étiquetées, on ré-applique la méthode d'adaptation de domaine (non-supervisée) pour étiqueter les données sources (en ignorant donc leurs vraies étiquettes). On utilise la qualité de ce ré-étiquetage des données sources (en comparaison avec les vraies étiquettes) comme mesure de qualité (du jeu d'hyper-paramètres). On teste donc si l'algorithme est cohérent en l'appliquant dans un sens puis dans l'autre. Cependant, il est à noter qu'un algorithme peut être très bon selon cette mesure mais très mauvais en pratique, par exemple s'il est parfait mais inverse toutes les étiquettes à chaque fois (dans un cas binaire, si on inverse deux fois, on retombe sur les bonnes étiquettes). De ce fait, pour être plus robuste, il nous est arrivé d'utiliser l'algorithme dont on veut optimiser les hyper-paramètres dans le sens source→cible mais un autre algorithme (imparfait mais plutôt stable, en fixant ses hyper-paramètres) dans le sens cible→source.

transport optimal et adaptation de domaine

Toute une partie des travaux présentés dans la section sec. IV.5 porte sur l'exploitation du transport optimal dans des tâches d'UDA. Le transport optimal est un formalisme pour obtenir une mesure de disparité entre deux distributions. Si l'on s'intéresse aux distributions de données sources et cibles, il constitue alors un outil naturel pour l'adaptation de domaine, même si son formalisme du transport optimal est bien plus large que la question de l'UDA ou même de celle de l'apprentissage par transfert.

IV.2 📦 Adapation de domaine par alignement de sous-espaces

stage sur l'alignement non-linéaire

La première contribution [58] en adaptation de domaine à laquelle j'ai participé à mon arrivée au laboratoire Hubert Curien s'est développée dans le contexte du stage de Rahaf Aljundi, co-encadrée avec Marc Sebban et Damien Muselet. Ce stage s'est extrêmement bien passé et Rahaf a ensuite fait une excellente thèse à KU Leuven (Belgique). Ces travaux visaient à étendre une méthode linéaire très simple mais efficace proposée précédemment au la-

boratoire, en lui permettant de capturer des aspects non-linéaires dans un contexte d'UDA.

subspace alignement (SA)

La méthode servant de base à notre contribution était SA (*Subspace Alignement*) [59]. SA cherche à aligner les deux domaines en apprenant une transformation linéaire. Cette dernière est appliquée sur les données sources, et un classifieur (par exemple un SVM) est appris sur ces données transformées pour être ensuite appliqué sur de nouvelles données (cibles). Pour apprendre cette transformation de manière robuste (et aussi sans qu'elle soit trivialement l'identité), les données sources et cibles sont projetées sur des sous-espaces source et cible. La solution au problème d'alignement s'obtient alors en effectuant deux PCA (principal components analysis) et en réalisant deux produits de matrices. Simple à comprendre et à programmer, cette méthode s'est également avérée donner de meilleurs résultats que des méthodes bien plus complexes.

inefficacité de la kernelisation de SA

Nous sommes partis du constat que cette méthode était probablement améliorable en la rendant non-linéaire. SA utilisant la PCA, une idée directe aurait pu consister à utiliser la version kernelisée de PCA, KPCA (*kernel-PCA*). Cependant, cette solution n'est pas adaptée. En effet, SA apprend une projection du sous-espace source vers le sous-espace cible, les deux étant des sous-espaces d'un même espace plus grand. La projection apprise par SA travaille dans l'espace complet et fait donc implicitement usage de la correspondance entre les dimensions entre les deux domaines (SA ne marche par exemple pas sur des domaines hétérogènes). Dans le cas de KPCA, les dimensions sources n'ont aucune correspondance avec les dimensions cibles, ce qui condamne la méthode à échouer. Pour donner un argument empirique, nous avons rapporté les résultats de SA sur deux KPCA, qui s'avèrent catastrophiques (proches du hasard, et bien pire que de ne pas adapter).

LSSA, SA kernelisé sur des landmarks communes

Nous avons donc proposé LSSA (*landmarks selection based subspace alignment*) [58], qui a aussi été intégré avec SA et exposé dans un chapitre de livre [57] (dans un livre sur l'adaptation de domaine [60]). Le principe de LSSA et de réaliser l'équivalent d'une KPCA mais en utilisant un ensemble de points commun entre les deux domaines. Un noyau gaussien est appliqué entre les points des ensembles source et cible d'un côté, et des points de référence appelés *landmarks* d'un autre côté. Même si prendre comme landmarks l'ensemble des points ou certains points aléatoirement est efficace, nous sommes allés plus loin en proposant une méthode de choix encore plus performante. Ce manuscrit utilise le mot *landmark* et le considère comme un nom féminin, du fait que la traduction mot-à-mot pourrait être *marque* (ou *référence*, ou *borne*).

améliorer le choix des landmarks

Celle-ci se base sur un critère de qualité de landmark. Ce critère mesure la similarité de distributions (entre les points sources et les points cibles) des valeurs du noyau évalué entre la landmark et un point. Plus précisément, pour mesurer la qualité d'un point candidat à être landmark, nous appliquons d'abord le noyau (par exemple gaussien) entre le point candidat et l'ensemble des points sources et cibles. Notre but est alors de mesurer la similarité entre la distribution des valeurs pour les points sources et celles pour les points cibles, cela permettant intuitivement de choisir des landmarks qui projettent les deux jeux de données de la même façon. Comme le noyau gaussien devient rapidement constant (décroissance rapide vers zéro), on cherche en fait des landmarks qui projettent bien les deux jeux de données, mais localement.

similarité entre distributions

Pour mesurer la similarité entre nuages de points discrets, nous avons décidé d'abord de réaliser une estimation de la densité de chaque groupe de points, sous forme d'une loi normale. Ce choix est probablement trop simpliste et peu adapté à ces données qui sont toutes positives avec beaucoup de valeurs en 0 mais c'est une estimation qui est robuste et permet des calculs en forme close. En effet, comme mesure de similarité entre deux distributions normales, nous avons pris l'intégral du produit de leurs densités, qui peut s'exprimer à partir des moyennes et variances des deux distributions normales. Cette similarité est alors renormalisée de façon à ce qu'une valeur de 1 soit obtenue si les distributions sont parfaitement alignées, de façon à faciliter l'établissement d'un seuil cohérent pour le choix des landmarks effectivement sélectionnées.

choix de la variance du noyau

En plus de choisir les landmarks (points), nous choisissons aussi la variance du noyau gaussien qui maximise le recouvrement pour chaque landmark. Il est intéressant de remarquer, a posteriori, que la mesure de recouvrement proposée résout (différemment) une problématique similaire à celle du transport optimal discret, qui est maintenant beaucoup utilisé en particulier pour l'adaptation de domaine (voir section IV.5). Nous aurions pu essayer une variété de mesures de similarité entre distributions (ou nuages de points) et explorer un choix de landmarks incrémental, conditionnellement aux landmarks déjà choisies, mais du fait de contraintes de temps liées à la durée du stage, ces travaux se sont limités à comparer le choix proposé à un tirage aléatoire des landmarks et au cas où la variance par landmark n'est pas choisie.

IV.3 Adaptation de domaine, apprentissage multi-tâche et profond

nouvelle couche pour la color constancy

Dans le cadre de la thèse de Damien Fourure, co-encadré au laboratoire avec l'équipe image, mais aussi avec Christian Wolf du LIRIS, nous avons réalisé différentes contributions autour des réseaux de neurones. Les premiers travaux visaient à prendre en main et mieux comprendre ce qui constituait

encore à l'époque de nouvelles approches de machine learning pour la vision par ordinateur. Nous avons travaillé sur la tâche d'estimation de couleur de l'éclairage (*color constancy*). Les méthodes sans apprentissage travaillaient avec des statistiques locales ou plus globales, sous forme de moyennes ou maximum. Nous avons généralisé les architectures existantes en proposant un *mixed pooling* [1] qui réalise en parallèle une agrégation par une norme ℓ_p et par le maximum (norme ℓ_∞). Le calcul utilisant ces agrégations pour prédire automatiquement l'éclairage pouvant alors être appris par descente de gradient.

étiquettes hétérogènes pour la segmentation sémantique

Le cœur de la thèse de Damien Fourure a été dans le domaine de la segmentation sémantique d'image, en proposant des nouvelles approches à base de réseaux de neurones. La tâche visait à affecter une classe à chaque pixel d'une image, par exemple parmi les classes voiture, arbre, cycliste, etc. À l'époque, la quantité de données étiquetées était relativement limitée et très hétérogène : en effet, comme l'annotation complète d'une image à l'échelle du pixel nécessite beaucoup d'attention et prend au minimum 30 minutes par image, quelques articles avaient annoté des images selon leurs besoins, donc avec des classes différentes et avec relativement peu de données et venant d'ensembles différents. Nous avons proposé une architecture et un processus d'apprentissage avec trois éléments principaux :

- une adaptation de domaine par inversion de gradient (*gradient reversal layer*) qui permet d'apprendre une représentation (de patch) d'image qui est invariante aux domaines, combinée avec un équilibrage des jeux de données (comme étudié plus tard dans sec. IV.4, le déséquilibre pouvant avoir un fort effet négatif sur l'adaptation de domaine),
- une fonction de perte sélective, la *selective loss* qui permet un apprentissage joint de plusieurs tâches en réalisant un *softmax* pour chaque jeu d'étiquettes (chaque tâche de classification),
- une couche permettant d'exploiter la corrélation entre les tâches et d'une certaine façon de mieux calibrer le classifieur multi-tâches.

publications et collaboration

Ces travaux ont été développés dans le domaine de la segmentation sémantique de scène de trafic (par exemple, pour les voitures autonomes) [2], [3], mais ont aussi établi un nouvel état de l'art dans le domaine de l'estimation de pose de la main, dans une collaboration [4] avec Natalia Neverova alors doctorante au LIRIS. Ces travaux ont été réalisés dans le contexte du projet ANR Solstice.

encadrement et candidat adapté

Malgré un nombre de co-encadrants très important (cinq au total dont trois ou quatre très fortement impliqués selon les thématiques) avec des expertises différentes, une réussite de la thèse de Damien Fourure est que nous ayons réussi à l'encadrer au bon niveau. Il a su développer une expertise

et a, en parallèle des nombreuses pistes que nous pouvions lui suggérer, proposé une nouvelle architecture de réseau en grille, *GridNet*, qui généralise les approches existantes (ResNET, UNet et plusieurs approches de segmentation sémantique). Malgré son côté générique, cette architecture, ainsi qu'une méthode de *dropout* spécifique (par bloc) se sont montrées très efficaces et battant de nombreuses approches [5], y compris celles utilisant un pré-apprentissage sur ImageNET.

IV.4 Adaptation de domaine et détection d'anomalies supervisée

une thèse CIFRE qui débouche sur des projets

Suite au stage de Kevin Bascol sur la fouille de motifs par auto-encodeurs, nous lui avons proposé, avec Élisa Fromont, de continuer en thèse. Nous avions une opportunité de financement de thèse CIFRE avec l'entreprise Bluecime qui travaille sur de la sécurité sur les télésièges et avait alors pris contact avec nous. Les discussions avec le dirigeant et les représentants techniques de l'entreprise ont révélé une volonté réelle de faire de la recherche avec un doctorant disponible à 100% au laboratoire. Ces travaux se sont donc positionnés dans le contexte de la thèse CIFRE de Kevin Bascol mais aussi du projet FUI MIVAO que nous avons déposé par la suite avec Bluecime pour le financement de postdoc d'Emmanuel Roux (et d'une thèse dans l'équipe image du laboratoire).

adaptation de domaine et données déséquilibrées

La thèse de Kevin Bascol s'est intéressée à des problématiques scientifiques d'adaptation de domaine dans un contexte de détection d'anomalies supervisée. Bien que ces travaux puissent d'une certaine façon rentrer dans le chapitre V, le cœur des travaux de thèse étaient plutôt en adaptation de domaine (et les contributions spécifiquement liées aux données déséquilibrées sont présentées dans l'autre chapitre). En effet, Bluecime avait déjà commencé à collecter une quantité d'anomalies relativement importante et la problématique principale était la variété des jeux de données (différents télésièges).

adaptation de domaine pour la détection supervisée de situation anomale

Une première contribution de la thèse [19] a été de montrer que, étant donnée la quantité de données déjà étiquetées pour cette tâche, l'apprentissage automatique pouvait remplacer l'approche programmée manuellement par l'entreprise. En effet, Bluecime avait déjà environ 4000 images de cas à risque (*unsafe*) pour réaliser des tests de non-régression sur leur système. À l'époque, les données étaient issues de 16 domaines (16 remontées mécaniques) avec des apparences et des points de vue bien différents, mais aussi des conditions de neige ou d'éclairage variables ⛶ .

l'apprentissage est plus précis que le système écrit manuellement

En utilisant un ensemble de non-régression pour développer un système, comme fait jusque là sans apprentissage automatique, il y a un risque de sur-spécialisation sur ces données de non-régression. C'est l'équivalent du sur-apprentissage mais réalisé par le processus de développement logiciel par l'humain. Nous avons appliqué un protocole expérimental rigoureux d'apprentissage et évalué les différentes situations (apprendre sur des données sources, apprendre sur des données sources et tester sur une nouvelle remontée, etc), sans ou avec adaptation de domaine. L'adaptation de domaine rapportée dans [19] est faite à l'aide d'un apprentissage adversaire avec un classifieur de domaine et une couche d'inversion de gradient, mais nous avons exploré par la suite différentes formulations de fonctions de perte adversaires. Nous avons pu montrer que l'adaptation de domaine était capitale lors du test sur une nouvelle remontée et apportait un petit gain si l'on a déjà des données de la remontée de test. Dans tous les cas, nous avons pu montrer que le système créé manuellement par l'entreprise était moins performant que le système que l'on pouvait apprendre.

pondérations des sources par proximité au domaine cible

Une seconde contribution autour de l'adaptation de domaine [22] a été de considérer un contexte dit multi-sources, bien adapté au cas d'application de Bluecime. La question scientifique posée fut celle du choix des domaines sources à utiliser pour éviter le transfert négatif, c'est-à-dire éviter d'utiliser un domaine source qui dégrade les performances sur le domaine cible d'intérêt. Comme cette tâche ne peut pas utiliser d'étiquettes dans le domaine cible, nous l'avons formulée en utilisant la proximité des données entre domaines (en ignorant les étiquettes). La méthode est modulaire : elle permet d'utiliser différentes distances entre domaines (par exemple une distance basée sur le transport optimal). Le principe est d'utiliser les distances entre le domaine cible et tous les domaines sources, de transformer ces distances en scores, puis de normaliser ces scores pour avoir un vecteur de probabilité. Le vecteur de probabilité est alors utilisé pour sélectionner les exemples utilisés dans l'apprentissage adversaire, servant ainsi de pondération des différents domaines sources.

taille des domaines et diversité

Dans ce contexte non-supervisé, certains domaines sources sont petits en nombre d'exemples et il peut être catastrophique de n'utiliser que ce genre de domaine : si de petits domaines sont proches du domaine cible, alors très peu d'exemples seront utilisés pour apprendre l'espace de représentation multidomaine. Nous avons donc introduit une mesure de variété sur le vecteur de probabilité qui prend en compte les tailles des domaines et qui approxime le nombre d'exemples différents utilisés lors de l'apprentissage adversaire. Cette mesure de variété est utilisée pour raffiner les poids utilisés au final dans l'apprentissage adversaire. Nos expériences, sur des jeux de données standard et sur les jeux de données de Bluecime, ont montré l'intérêt de l'approche.

intégration rapide

Les travaux de Kevin Bascol ont été rapidement intégrés dans le système de Bluecime et mis en production pendant la période de la thèse. Suite à sa thèse et un court postdoc sur le projet MIVAO, Kevin a été embauché en CDI dans son entreprise.

attention aux variations de déséquilibre

Il me semble utile de mentionner un point clé qui nous est apparu au fil des expériences mais que nous n'avons pas pu intégrer de manière satisfaisante dans nos contributions. Par défaut, les méthodes d'adaptation basées sur un rapprochement de distributions (comme les approches adversaires qui sont reliées au transport optimal) sont sensibles à la différence de proportion des classes d'un domaine à l'autre. En effet, s'il y a plus d'éléments d'une classe donnée dans le domaine source, alors une partie de ces éléments seront projetés/associés à des éléments cibles de l'autre classe. Nous avons pu observer qu'en cas de variation de déséquilibre, les méthodes d'adaptation pouvaient avoir un effet négatif. D'une certaine façon (mais de manière imparfaite), le choix des domaines peut réduire ce problème.

IV.5 Transport optimal et adaptation de domaine

cf manuscrit de Tanguy Kerdoncuff

La thèse de Tanguy Kerdoncuff, que j'ai co-encadrée avec Marc Sebban, portait sur l'adaptation de domaine, mais plus particulièrement sur les méthodes de transport optimal. Cette thèse a mené à des contributions aussi bien théoriques, qu'algorithmiques et pratiques. Son manuscrit de thèse [61] est d'ailleurs un bon point d'entrée pour retrouver plus de détails généraux sur le transport optimal, par rapport à ce document.

IV.5.1 Problème du transport optimal

transport optimal, formulation discrète

Le problème du transport optimal est un problème d'alignement de distributions de probabilité. Dans sa version discrète, il prend en entrée :

- deux ensembles de points, que nous pouvons de manière pratique appeler source $A = \{A_i\}_{i=1}^{\#A}$ et cible $B = \{B_j\}_{j=1}^{\#B}$, respectivement,
- des masses associées à tous ces points, $a = \{a_i\}$ et $b = \{b_j\}$,
- une matrice $\#A \times \#B$ de coût $C = \{C_{ij}\}$ (coût sous-jacent ou « métrique de sol », *ground metric*) qui donne le coût associé au transport d'une unité de masse d'un endroit source A_i à un endroit cible B_j.

un problème linéaire avec contraintes

Le problème du transport optimal revient à trouver le coût global minimal pour transformer la distribution source en la distribution cible, ou plus formellement $\min_{T \in \Pi_{a,b}} \sum_{i,j} T_{ij} C_{ij}$.

Le but est de trouver le plan de transport T (et c'est explicitement le but si on remplace le min par un arg min) qui minimise le coût global. Le plan de transport est une matrice qui explicite quelle masse il faut envoyer d'un endroit i à un endroit j. Pour être admissible, ce plan de transport doit être dans $\Pi_{a,b}$, l'ensemble des matrices bi-stochastiques de marginales a et b, c'est-à-dire que T doit être une matrice avec des valeurs non-négatives et doit vérifier $\forall i, \sum_j T_{ij} = a_i$ et $\forall j, \sum_i T_{ij} = b_j$. Ce problème est historiquement utile pour effectuer l'acheminement optimal de ressources (mines vers usines, redéploiement de troupes, transport de remblai, etc). Le transport optimal s'avère être un formalisme très adapté en apprentissage automatique, pour mettre en correspondance des distributions de données. Le problème peut être directement formulé dans un espace continu, voire entre distributions discrètes (sommes de Diracs) et continues.

IV.5.2 Transport optimal et apprentissage de métrique

alignement de sous espaces et apprentissage de métrique

La première contribution de la thèse de Tanguy Kerdoncuff a porté sur une approche mêlant transport optimal et apprentissage de métrique pour l'adaptation de domaine non supervisée [28], appelée MLOT (*metric learning in optimal transport*). Nous avons conçu cette approche en nous positionnant dans un contexte de transport optimal tout en capitalisant sur les connaissances venant entre autres des méthodes d'alignement de sous-espaces (SA et LSSA, voir sec. IV.2).

réductions différentes

En effet, il s'avère capital dans SA de réaliser des extractions de sous espaces pour que l'alignement (la projection) ait un effet. De plus, ces sous-espaces doivent être spécifiques aux données sources d'un côté et aux cibles de l'autre.

PCA et transport optimal

Ces travaux font, d'une part, un lien entre PCA et la minimisation de la distance de Wasserstein, plus précisément sous la forme d'un barycentre au sens de la Wasserstein-2 (qui travaille avec les carrés de la distance sous-jacente). Ceci est aligné avec le fait que la PCA minimise la variance résiduelle et donc des distances au carré [ ] . L'approche proposée réalise donc une réduction de dimension dans l'espace cible, et une réduction de dimension dans l'espace source mais celle-ci est guidée par les étiquettes disponibles. La divergence entre les ensembles de points dans les espaces projetés est mesurée par transport optimal. Cette formulation a plusieurs avantages :

- Elle permet de généraliser la méthode SA (*Subspace Alignement*) en intégrant l'apprentissage d'une métrique grâce aux étiquettes disponibles dans l'espace source.
- Elle utilise une mesure de transport optimal pour évaluer la divergence entre les domaines (projetés), qui est plus robuste et donne de meilleurs

généralisation de SA

stabilité du transport

informations de gradient que les formulations de type *gradient reversal layer*.

optimisable

- Elle propose ainsi une fonction à minimiser (par exemple par descente de gradient) qui permet d'optimiser la réduction de dimension tout en y intégrant potentiellement des régularisations proposées dans d'autres méthodes.
- Elle nous permet de dériver des garanties en généralisation de l'approche par un traitement uniforme de toute la chaîne de transformation : il existe des bornes (en termes de Wasserstein) entre une distribution et un échantillon de cette distribution, nous avons de plus établi un lien entre la réduction de dimension et la distance de Wasserstein, et l'approche minimise explicitement la distance de Wasserstein entre les sources projetées et cibles projetés [ ] .

avec garanties

IV.5.3 Transport optimal hétérogène : problème de Gromov-Wasserstein (GW)

au delà de Wasserstein

La suite des travaux de thèse de Tanguy Kerdoncuff se sont intéressés principalement aux généralisations de la formulation du transport optimal et au passage à l'échelle de celles-ci. Une limitation de la formulation du transport optimal est qu'elle repose sur une fonction de coût qui compare un endroit source et un endroit cible. Ainsi, le transport optimal est souvent utilisé dans un cadre où les espaces source et cible sont identiques (ou des sous-espaces d'un même espace) et une distance ou métrique associée à cet espace est utilisée comme fonction de coût sous-jacente. Ceci permet aussi d'avoir de bonnes propriétés sur la mesure de Wasserstein qui est alors une distance entre distributions.

GW, comparaison de paires

Un nouveau formalisme, appelé *Gromov-Wasserstein* s'intéresse à généraliser le transport optimal à des espaces de natures différentes, sans nécessiter de définir une fonction de coût entre ces espaces. La solution repose toujours sur un transport des points sources sur les points cibles, mais l'idée est de ne comparer que des paires sources avec des paires cibles.

formulation OT : rappel

Le problème de transport optimal classique repose sur une fonction de coût entre un point source et un point cible. C'est cette fonction de coût qui rempli la matrice C dans la version discrète.

comparaison de descripteurs de paires

Le problème de Gromov-Wasserstein repose de son côté sur une fonction de coût $\mathcal{L}$ qui compare plutôt une paire source avec une paire cible. Le « descripteur » d'une paire est souvent considéré comme étant un coût (ou une distance) dans les travaux existants, et sera noté ici C^A pour les sources et C^B pour les cibles.

decripteur non-scalaire de paire

Pour ne pas limiter l'applicabilité de ces approches, il est important de noter

que ce « descripteur de paire » peut en fait être quelconque et potentiellement dans un espace à plusieurs dimensions. Nous avons reformulé avec cette vision un problème de distillation de réseau de neurones récurrent en un problème de barycentre au sens d'une distance de *fused gromov-wassertein* .

Dans sa version discrète, le problème de Gromov-Wasserstein peut être formulé ainsi :

$$GW(\mathcal{L}, C^A, C^B) = \min_{T \in \Pi_{a,b}} \sum_{i,i',j,j'} T_{ij} T_{i'j'} \mathcal{L}\left(C^A_{ii'}, C^B_{jj'}\right)$$

où les indices i et i' sont sur l'espace source A, et j, j' sur le cible B, l'unique plan de transport T associant bien les points de A avec les points de B. Le problème peut aussi être reformulé conceptuellement par un pré-calcul des coûts $\mathcal{L}$ dans un tenseur à 4 dimensions L :

$$GW(L) = \min_{T \in \Pi_{a,b}} \sum_{i,i',j,j'} T_{ij} T_{i'j'} L_{ii'jj'}$$

somme sur les paires de paires

Grâce à son raisonnement par paires, GW est un outil très adapté pour comparer deux graphes avec des nœuds différents en utilisant par exemple pour C^A la matrice d'adjacence du graphe source (le graphe pouvant être pondéré).

Il est possible de reformuler plusieurs méthodes de clustering ou de réduction de dimension, notablement *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP) comme un problème de barycentre au sens de Gromov-Wassertein, en utilisant les bonnes distances et fonctions de pertes. Ceci ouvre la possibilité de formuler ce type de problèmes dans un contexte de transport optimal, et d'utiliser toutes les connaissances autour du transport et de ses variantes, des algorithmes de résolutions, etc.

barycentre comme formulation flexible

IV.5.4 Transport optimal : résolution et complexité

Le problème OT est un problème linéaire mais difficile car de grande taille et sous contraintes. Il est linéaire en les entrées du plan de transport T puisque l'on minimise un produit scalaire $< T, C >$. Pour simplifier, nous supposons qu'il y a N points dans chacun des deux ensembles source et cible. Il y a alors N^2 entrées dans le plan de transport T (et dans la matrice de coût C). Nous nous intéressons d'abord à la complexité (en temps) du problème de transport OT classique.

problème linéaire en N^2

Cette complexité est au minimum de $\mathcal{O}(N \log(N))$ dans le cas à une dimension. Intuitivement, **en 1D** et si la fonction de coût est monotone, on peut ignorer les valeurs de coût et il suffit de trier les points puis de faire une affectation glissante.

Pour des espaces à **plus d'une dimension**, la solution générale est celle

$N \log(N)$ en 1D, sans matérialiser la matrice C, $N^3 \log(N)$ sinon

d'un problème de flot de coût minimal, résolu par un algorithme de type *simplex* avec une complexité en $\mathcal{O}(N^3 \log(N))$ dans le pire cas.

régularisation entropique, Sinkhorn, N^2

L'algorithme Sinkhorn est souvent utilisé en pratique de part son meilleur passage à l'échelle, son côté itératif et sa formulation différentiable. C'est un algorithme approché (et qui résout le problème augmenté d'une régularisation entropique sur le plan de transport) et itératif. La régularisation entropique vise à maximiser l'entropie de T, c'est-à-dire à le rendre le moins parcimonieux possible (avec un coefficient, habituellement noté ε, pour contrôler le poids de cette régularisation). Même si cela va à l'encontre d'une volonté qui pourrait être d'avoir une matrice de transport creuse (avec les possibles avantage en termes de calculs), c'est cela qui rend le problème plus « simple » dans le sens où les solutions ne sont plus au bord du domaine. En ignorant le nombre d'itérations (que l'on peut lier à la quantité de régularisation entropique ajoutée), on obtient une complexité de $\mathcal{O}(N^2)$.

régularisation = a priori

Il est intéressant de noter (pour l'interprétation et la généralisation des approches) que la régularisation entropique peut aussi être vue dans un contexte probabiliste comme un a priori, et aussi sous forme d'une méthode de descente de gradient proximale au sens de la divergence de Kullback-Leibler.

Complexité de GW

Le problème de Gromov-Wassertein (GW) est un problème quadratique en T (qui lui est déjà de taille N^2). Dans le cas le plus général, la complexité nécessaire à calculer tous les termes de la somme du problème de GW est déjà en $\mathcal{O}(N^4)$. Face à cette complexité, des approches itératives par descente de gradient sont utilisées. Une régularisation entropique est de plus utilisée pour pouvoir utiliser l'algorithme de Sinkhorn (comme sous procédure, à chaque itération de l'algorithme global). Le problème régularisé est noté EGW (*Entropic Gromov Wasserstein*) et vise à minimiser le problème de GW et maximiser l'entropie du plan de transport T.

GW en itérant sur OT

Intuitivement, le principe itératif de résolution de GW est d'initialiser la résolution avec un plan de transport (par exemple un plan uniforme, c'est-à-dire qui envoie chaque point source sur l'ensemble des points cibles). Ensuite, chaque itération va remplacer un des T dans le problème de GW par le plan de transport courant et optimiser sur l'autre T, par la résolution d'un problème de transport classique. Cela correspond en fait à une procédure de descente de gradient où chaque pas demande de résoudre un problème de transport optimal. En effet, en fixant un des plans à T et en optimisant sur le second T', on peut ré-écrire le terme à minimiser :

$$\sum_{i,i',j,j'} T_{ij} T'_{i'j'} L_{ii'jj'} = \sum_{i',j'} T'_{i'j'} \sum_{i,j} T_{ij} L_{ii'jj'} = < T', \Lambda >$$

calcul de matrice de coût, décomposable ou non

La matrice de coût du problème de transport intermédiaire est le produit (partiel) entre L et l'estimation courante du plan de transport T, et est notée $\Lambda_{i',j'} = \sum_{i,j} T_{ij} L_{ii'jj'}$ ou $\Lambda = \sum_{i,j} T_{ij} L_{i.j.}$ qui fait apparaître explicitement une somme de matrices (utile pour la suite). Dans le cas général, le calcul de Λ a déjà une complexité de $\mathcal{O}(N^4)$, ce qui peut déjà être bloquant pour les applications. Dans les cas où $\mathcal{L}(C^A, C^B)$ peut être décomposée (sous la forme de $f_1(C^A)+f_2(C^B)-h_1(C^A)h_1(C^B)$), la matrice de coût intermédiaire peut être calculée en $\mathcal{O}(N^3)$.

Approches de type « *sliced* »

Étant donnée la complexité du problème de transport (OT), une approche par projections aléatoires (*sliced-Wasserstein* [62]) permet de se ramener en un problème 1D et d'utiliser un résolution en $\mathcal{O}(N \log(N)))$. Cette approche *sliced*, bien que très rapide a plusieurs limites : elle ne calcule pas (ni n'approxime) la distance de Wasserstein mais définit un autre problème, elle ne fournit pas explicitement de plan de transport et le plan que l'on pourrait implicitement en dériver est très différent du plan de transport optimal.

IV.5.5 📦 Passage à l'échelle de Gromov-Wassertein

approche stochastique PoGroW, N^2 voir $N \log(N)$

Une des contributions de la thèse de Tanguy Kerdoncuff a été de proposer des algorithmes stochastiques de plus faible complexité algorithmique que les approches existantes, en particulier sur les problèmes de type Gromov-Wassertein. Cela est possible grâce à un algorithme stochastique (Frank-Wolfe stochastique) qui permet d'éviter le temps de calcul en $\mathcal{O}(N^4)$ de la matrice Λ, nécessaire à chaque itération. L'avantage de cette approche est qu'elle travaille bien avec (une approximation de) la vraie distance de GW, se prête à la dérivation de garanties de convergences et à une complexité d'itération en $\mathcal{O}(N^2)$. Cette approche est nommée *Sampled Gromov-Wassertein*, SaGroW, [29]. Un cas particulier de l'approche, nommé *pointwise Gromov-Wassertein*, PoGroW, peut même donner lieu à une complexité d'itération en $\mathcal{O}(N \log(N))$ grâce au passage à un problème intermédiaire de transport en une dimension. Nous avons aussi proposé une variante avec une régularisation utilisant la divergence de Kullback-Leibler, donnant lieu à une descente de gradient au sens de la divergence de Kullback-Leibler et l'utilisation de l'algorithme de Sinkhorn.

faire apparaître une espérance à échantilloner

Intuitivement, l'approche stochastique vient de la formule exposée précédemment $\Lambda = \sum_{i,j} T_{ij} L_{i.j.}$, qui fait apparaître une somme de matrices. Cette somme est en particulier pondérée par T, qui peut être vu comme une distribution, puisque T somme à 1. La matrice de coût intermédiaire Λ peut donc être vue comme l'espérance d'une variable aléatoire à valeur matricielle (dont le domaine est formé des matrices $L_{i.j.}$ avec probabilités T_{ij}). C'est la clé de l'approche stochastique et nous remplaçons cette espérance exacte

(dont le calcul est en $\mathcal{O}(N^4)$) par un échantillonage avec une liberté sur le nombre d'échantillons M (pour un coût de $\mathcal{O}(MN^2)$). Le choix de M permet de contrôler le compromis entre précision et vitesse. De plus, comme T est relativement parcimonieux, l'espérance peut intuitivement être bien approximée même avec un nombre d'échantillons limité. Une approche stochastique peut aussi avoir des propriétés intéressantes de robustesse, comme la descente de gradient stochastique pour les réseaux profonds (mais dans lesquels la fonction à optimiser est beaucoup plus irrégulière que dans le cas du transport optimal). Un des avantages de la méthode stochastique que nous avons proposée est qu'elle n'est pas contrainte par la forme de la fonction de coût (contrairement au cas particulier en $\mathcal{O}(N^3)$).

évaluation de la valeur de GW

Nos approches stochastiques travaillent par itération sur la valeur du plan de transport T. Dans ce but, elles ne calculent jamais la distance GW, qui a, dans l'absolu, un coût de $\mathcal{O}(N^4)$. Un potentiel défaut serait donc que nous n'avons pas accès à cette distance pouvant être utile dans certaines applications. Cependant, comme le plan de transport est parcimonieux (contrairement à d'autres approches), nous avons pu proposer un algorithme stochastique proposant une estimation non biaisée de GW. Nous avons évalué cet estimateur et montré qu'il estime extrêmement bien la valeur de GW avec un coût de $\mathcal{O}(N^2)$ et n'est donc pas plus coûteux que SaGroW (mais plus que PoGrow).

IV.5.6 Au delà du transport et de Gromov-Wassertein

extensions existantes de OT

Une autre contribution majeure de la thèse de Tanguy Kerdoncuff est une extension générale des problèmes de transport optimal. Nous avons vu que le problème OT classique travaille généralement sur des données dans un même espace (ou avec une fonction de coût inter-espaces) et cherche un plan de transport entre les points sources et les points cibles (dans sa version discrète). Nous avons aussi vu que le problème de GW, qui cherche aussi un plan de transport, travaille sur des espaces différents en utilisant des « distances » (descripteurs) entre paires. Il existe aussi d'autres formulations comme *Fused Gromov-Wasserstein* qui combine les deux (OT et GW) mais qui a la même contrainte que le problème OT classique (même espace). Une autre variante est le problème de *Co-Optimal Transport*, CoOT, qui travaille sur des espaces différents mais cherche alors deux plans de transport. Le premier plan est classique et met en relation les points sources avec les points cibles. Le second plan met en relation les features sources avec les features cibles, sur le même principe que pour les points. Cette méthode a notamment montré son intérêt en co-clustering.

généralisation *OTT*

Dans la thèse de Tanguy Kerdoncuff, nous avons généralisé toutes ces ap-

proches dans un cadre unique que nous avons appelé *Optimal Tensor Transport* (OTT, [30]). Bien que cette approche permette des versions *fused* en mélangeant autant de sous-problèmes que voulu, nous ne décrivons ici que la version de base (sans fusion de sous-problèmes). Avec OTT, nous avons défini un cadre incluant les problèmes précédents OT, GW, Co-OT, mais aussi une infinité d'autres cas. De manière générale (nous détaillons les cas classiques juste après), une instance de problème d'OTT discret est définie par :

- un tenseur source, avec une dimension de features et autant d'autres dimensions que voulu,
- un tenseur cible, avec le même nombre de dimensions mais pas forcément la même taille (nombre d'éléments),
- une spécification de combien de plans de transports doivent être trouvés, et de comment ils s'appliquent aux données,
- une fonction de coût pour comparer un vecteur de features sources avec un vecteur de features cibles.

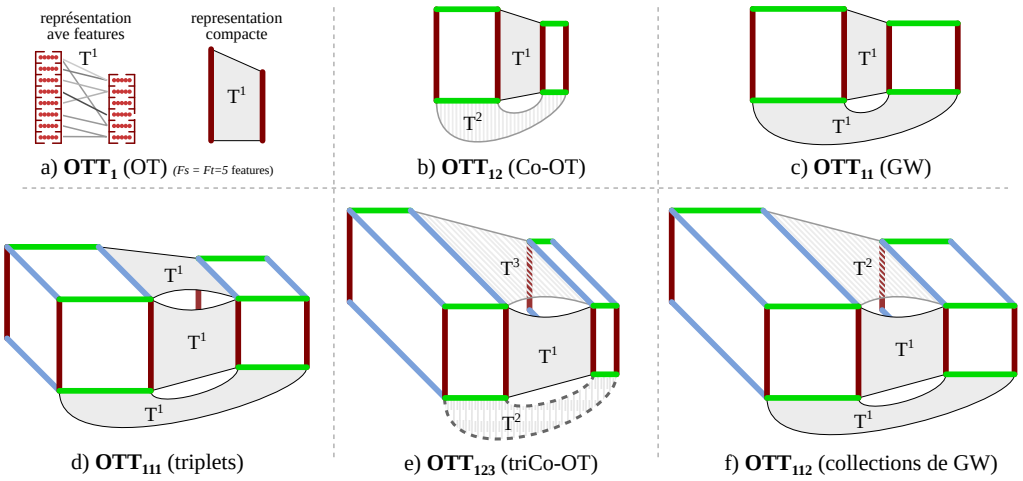


Figure IV.1: Diverses variantes de OTT avec, pour chaque cas, deux jeux de données et les plans de transports recherchés.

Le cas du transport optimal classique prend en entrée une matrice $N_{source} \times F$ et une $N_{cible} \times F$. Le problème OT est alors noté OTT_1 : il y a un seul indice, car seulement la première dimension des tenseurs est mise en correspondance / transportée, et cet indice est 1 car c'est le premier (et unique) plan de transport qui est utilisé pour cette dimension. Ici, F est le nombre de features et la fonction de coût est la métrique sous-jacente que l'on utilise dans le transport optimal. Plus généralement, F pourrait aussi être une dimension vide (un seul scalaire pour décrire un point) ou plusieurs dimensions, voir même être différent entre source et cible. La contrainte est que la fonction de coût soit cohérente avec ces dimensions et puisse donner une valeur de coût à chaque combinaison source/cible.

OT = OTT_1

GW = OTT_{11}

Le cas de Gromov-Wasserstein (GW) prend en entrée une matrice $N_{source} \times N_{source}$ et une $N_{cible} \times N_{cible}$ (on pourrait ajouter une dimension de features si l'on a un descripteur par paire au lieu d'un scalaire). Le problème de GW est noté OTT_{11} car il utilise, à la fois pour la première et la seconde dimension le même (unique) plan de transport. Cette vision générale permet bien de ré-imaginer un GW avec un descripteur non-scalaire pour les paires.

Co-OT = OTT_{12}

Dans le cas du *Co-Optimal Transport*, il faut aussi le voir avec un descripteur *scalaire* et ceci malgré la présence de « features » dans les matrices données qui sont $N_{source} \times F_{source}$ et $N_{cible} \times F_{cible}$. Le problème de Co-OT est noté OTT_{12} car la première dimension utilise un premier plan de transport et la seconde dimension (celle des features) utilise un second plan de transport. Nous voyons déjà une extension directe de Co-OT qui utiliserait un descripteur non-scalaire pour les entrées, c'est-à-dire des ensembles de points à aligner, avec des ensembles de features aussi à aligner qui elles seraient multidimensionnelles (par exemple des groupes de features).

OTT_*

Toutes les variantes d'OTT peuvent être imaginées et utiles pour des cas d'utilisation existants ou à venir comme illustré dans la figure IV.1, . Par exemple OTT_{111} est une mise en correspondance de triplets, OTT_{112} est un GW mais où les descripteurs de paires sont des vecteurs dans des espaces différents (entre source et cible) et qui nécessitent d'être alignés comme dans Co-OT. OTT_{123} est le cas d'alignement de points qui sont eux mêmes des tableaux qui doivent être alignés comme dans Co-OT. En imaginant toutes les variantes de OTT, combinées à la variété de scénarios possibles avec la fonction de coût et les dimensions non-alignées (dimensions de features), les perspectives d'utilisation sont importantes. Il est aussi possible de faire des versions *fused* de OTT, en combinant autant de sous-problèmes que voulus et disponibles.

complexité extrêmes et approches stochastiques

Un des verrous d'OTT provient de sa résolution. La complexité croît exponentiellement (sur le modèle de GW) avec le nombre de dimensions (de l'ordre de $\mathcal{O}(N^{2D})$ s'il y a D dimensions à aligner avec chacun des N éléments en source et cible). Dans ce contexte de complexité extrême, les approches stochastiques développées pour GW (voir section IV.5.5) précédemment deviennent indispensables. Nous avons pu faire évoluer les approches stochastiques et les appliquer à OTT. Dans tous les cas, ces algorithmes gardent une complexité importante. Ils sont aussi relativement peu contraints, comme Co-OT ou GW le montrent déjà (d'où entre autres l'intérêt des approches *fused*). Il peut être utile de les appliquer dans un contexte semi-supervisé, dans le sens où une partie du plan de transport est donnée ou qu'une forme

d'a priori est connue. Ceci est particulièrement utile pour l'adaptation de domaine partiellement supervisée.

IV.5.7 Transport optimal robuste

transport optimal en pire cas sur la fonction de coût

En parallèle de ses sujets principaux de thèse, Tanguy Kerdoncuff a aussi collaboré avec un autre doctorant, Sofien Dhouib, sur une formulation du problème de transport optimal robuste [27]. Le principe est de chercher un plan de transport qui minimise le coût de transport, mais au lieu de le faire avec une fonction de coût sous-jacente donnée, le problème est résolu en pire cas sur les fonctions de coût d'une famille donnée. On a donc une formulation de type minimax ($\min_{plan} \max_{f_{cout}}$). Cette formulation du problème de transport optimal robuste et l'algorithme proposé sont des contributions importantes, et peuvent servir à obtenir des solutions à certains problèmes, par exemple en présence de bruit. Dans cette collaboration, nous avons en particulier apporté la notion de stabilité de matrice de coût. Le principe est de mesurer, pour une matrice/fonction de coût donnée, la variation maximale de coût de transport que l'on peut observer si l'on varie la matrice de coût localement (ajout de bruit à la matrice de coût). Nous avons pu expliciter un lien entre cette robustesse (de la matrice de coût) et la résistance au bruit de la distance de Wasserstein (coût de transport) : les matrices de coût plus robustes (du point de vue de notre nouvelle mesure) donnent des problèmes dont la solution est moins sensible au bruit.

Chapitre V

Apprentissage avec données déséquilibrées

Publications: [63] [64] [65] [66] [21] [67] [26] [20].

Projets : TADALoT, MIVAO.

Codes et liens divers

- Code d'exemple pour γ kNN.
  https://github.com/twitwi/gamma-kNN
- Code pour les expériences de MLFP, publié à la conférence IJCAI 2020.
  https://github.com/twitwi/MLFP

---

classification supervisée avec données déséquilibrées

Ce chapitre présente des travaux d'apprentissage automatique avec des données déséquilibrées, dans le sens où certaines classes sont plus présentes que d'autres. Ceci est par exemple le cas dans un contexte de détection de fraudes ou d'anomalies. Le problème est ici formulé sous la forme d'une tâche de classification supervisée, avec moins d'exemples de la classe d'intérêt (fraude, anomalie, …). Les terminologies utilisées peuvent être parfois déroutantes : la classe d'intérêt est souvent appelée la classe *positive* mais c'est aussi la classe *minoritaire*.

continuum classification déséquilibrée, détection d'anomalies ?

Le taux de déséquilibre (*imbalanced ratio*, IR) entre les classes peut être assez variable d'une situation à l'autre, allant typiquement d'un point d'intérêt pour deux points normaux, à un point d'intérêt pour plusieurs centaines ou milliers de points normaux. Les travaux présentés dans le chapitre III sur

les approches non-supervisées peuvent servir pour la détection d'anomalies ou plutôt d'événements' inhabituels. Plus le déséquilibre est grand, et donc plus le nombre d'exemples positifs est petit, plus la détection supervisée doit se rapprocher des approches non supervisées pour espérer généraliser correctement. En effet, s'il y a peu d'exemples positifs, il est probable que l'ensemble d'entraînement ne contienne pas certains types d'anomalies et qu'il faille donc détecter les *outliers*.

V.1 Mesure de qualité en détection déséquilibrée + apprentissage pondéré

anomalies, événements, alertes, alarmes

Pour fixer un peu la terminologie utilisée, prenons un exemple dans lequel on veut détecter des situations dangereuses, par exemple sur un télésiège. On s'intéresse donc à détecter un type *d'événements*, plutôt rare par rapport à l'ensemble des données. Sauf précision contraire, on se retrouve dans la même situation lorsque l'on veut détecter des *fraudes* ou des *anomalies*, dans un contexte supervisé. Une fois un système développé et/ou appris, il permet de prédire la présence de l'événement d'intérêt. On appelle *alarme* ou *alerte* (ou *détection*) les situations pour lesquelles le système prédit la présence de l'événement d'intérêt. Le système peut faire deux types d'erreurs : une fausse alarme (ou faux positif) qui est une alarme alors que l'événement n'a pas eu lieu, et une détection manquée (ou faux négatif) qui correspond à un événement pour lequel le système n'a pas levé d'alarme.

exactitude = accuracy = perte 0-1 séparable

Les tâches de classification sont généralement évaluées en termes d'exactitude (*accuracy*, à distinguer du terme de *précision* qui correspond à autre chose) qui est la proportion de points bien classifiés. De nombreux algorithmes optimisent cette mesure, c'est-à-dire la somme sur tous les points de la fonction de perte 0-1 (qui vaut 0 si le point est bien classifié, et 1 sinon). Pour pouvoir réaliser de l'optimisation continue, la plupart des approches utilisent d'une part un modèle qui renvoie une valeur continue et d'autre part des mesures de qualité continues, basées sur un substitut de la fonction de perte 0-1, comme la *hinge loss* ou la différence au carré. Une des propriétés de l'exactitude, qui la rend pratique, est que la fonction à optimiser est séparable : la fonction se décompose comme une somme sur l'ensemble des points. C'est le cas non-séparable qui motive une partie des travaux décrits ci-dessous.

limites de l'exactitude

Dans un contexte de classification déséquilibrée, l'exactitude peut être inadaptée car elle donne une plus grande importance aux points négatifs du fait qu'ils sont plus nombreux. Pour s'en convaincre, on peut imaginer un jeu de données relativement simple comme celui de la figure V.1, en deux dimen-

sions avec une famille de classifieurs linéaires. En maximisant l'exactitude, on obtient un classifieur non-informatif qui va prédire que tout est de classe normale. En effet, s'il y a 10 points dans la zone anormale (et 990 points normaux), l'exactitude de ce classifieur, optimal et inutile, sera de 99%. De son coté, un classifieur linéaire classifiant bien les 10 points anormaux, serait faux sur environ 40 points normaux et aurait donc une exactitude de 96% mais sera plus « utile ».

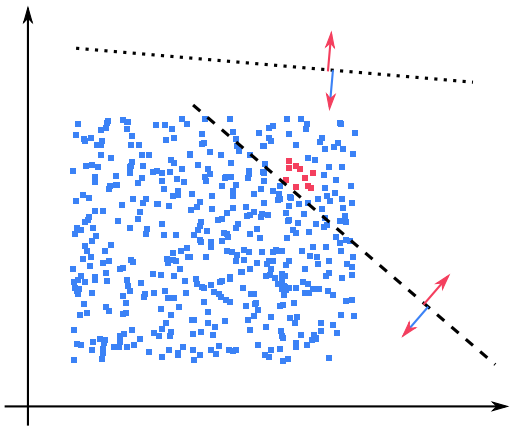


Figure V.1: Illustration d'un jeu de données déséquilibré, avec deux séparateurs linéaires. En haut en pointillés, le séparateur optimal en terme d'exactitude… mais inutile. En diagonale avec des tirets, un séparateur plus utile (mais non optimal).

classification déséquilibrée et ré-échantillonage

Partant du constat que l'exactitude est problématique et limitée pour les données déséquilibrées, plusieurs familles d'approches continuent à utiliser l'exactitude mais en essayant de ré-équilibrer les classes : les approches de sous-échantillonnage font une sélection des points négatifs pour réduire leur nombre, les approches de sur-échantillonnage répliquent des points positifs pour augmenter leur nombre ou génèrent de nouveaux points en se basant sur les points positifs. Ces familles d'approches contiennent de nombreuses variantes et elle ne sont pas exclusives, elles peuvent très bien être combinées entre elles ou avec les méthodes qui seront présentées dans ce chapitre.

apprentissage pondéré = ré-échantillonage virtuel

Le sous-échantillonnage a le défaut de ne pas pouvoir prendre en compte toute l'information (un défi est d'ailleurs de choisir des points plus représentatifs qu'une sélection aléatoire). De son côté, le sur-échantillonnage pose le problème qu'il augmente la taille du jeu de données (et donc le temps de calcul) sans ajouter d'information. Il risque aussi de créer des cas pathologiques, par exemple si on veut ensuite appliquer une forme de validation croisée (ou un modèle qui intègre implicitement une procédure comparable). La classification pondérée (ou apprentissage pondéré) considère un problème où il faut minimiser une exactitude pondérée, dans laquelle chaque classe (ou parfois

chaque point) a un poids. Intuitivement, l'apprentissage pondéré réalise une sorte de sur-échantillonage virtuel qui répliquerait les points de plus grand poids, mais ce sur-échantillonage est fait de manière analytique.

exactitude pondérée : intérêt et limites

L'exactitude pondérée est une bonne solution s'il est possible de donner des poids aux classes. En particulier, elle permet de donner des poids différents aux faux négatifs et aux faux positifs . Cependant, dans beaucoup de cas, il est difficile de fixer ces poids de manière pertinente. À cela s'ajoute que, dans les contextes applicatifs, l'exactitude pondérée peut être utile à optimiser, mais sa valeur n'a généralement pas de signification. Par exemple, on peut exprimer un compromis en disant qu'un type d'erreur (par exemple rater une situation dangereuse, c'est-à-dire un faux négatif) « coûte » 100 alors que l'autre type d'erreur (classer comme dangereuse une situation normale) ne coûte que 1. Cependant, si on obtient (sur un jeu de données) un coût total de 42000, cela ne donne pas beaucoup d'information : avons-nous raté 420 situations dangereuses ? ou avons-nous fait 42000 fausses alertes ? ou une combinaison de tout ça ?

précision, rappel

Les limitations de la mesure d'exactitude font que l'on utilise généralement des mesures spécialisées. Bien que peu concise, on peut mentionner la matrice de confusion qui reporte les nombres de classifications correctes des deux classes (TP, *true positive*, les alarmes correctes et TN, *true negative*, pour les cas normaux bien classés) et les nombres d'erreurs (FP, *false positive*, les fausses alarmes et FN, *false negative*, les anomalies manquées). D'autres mesures, relatives et plus synthétiques, sont aussi utilisées, telle que la *précision*, le *rappel* (*recall*) ou la F_β. La *précision* quantifie à quel point on peut faire confiance au système quand il déclenche une alerte : c'est la proportion des alarmes déclenchées par le système qui sont correctes, $\frac{TP}{TP+FP}$. Le *rappel* quantifie à quel point le système couvre les anomalies : c'est la proportion des anomalies qui sont détectées par le système, $\frac{TP}{TP+FN}$.

mesure F_β, score F_1

La mesure F_β est souvent utilisée pour synthétiser la précision et le rappel en une seule valeur, sous la forme d'un compromis entre précision et rappel. C'est en fait une famille de mesures (paramétrée par β) qui inclut par exemple la mesure F (avec $\beta = 1$, aussi appelée score F_1). La mesure F_β est une forme de moyenne harmonique (moyenne dans l'espace des inverses) pondérée par β^2 :

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot rappel}{\beta^2 \cdot precision + recall}.$$

F_β est non-séparable

Lorsqu'on veut l'optimiser, une difficulté d'une mesure de performance comme la mesure F_β est qu'elle est *non-séparable*. Cela veut dire qu'elle

ne se décompose pas comme une somme (ou moyenne) sur l'ensemble des points d'une fonction de perte associée à chaque point. Implicitement, mal classer un point implique une pénalité qui dépend d'à quel point les autres points sont bien classés. Il n'est donc plus possible de voir l'objectif comme une espérance sur la distribution des données (empirique ou théorique). Un objectif non-séparable rend donc moins aisée l'optimisation et les raisonnements en termes de garanties théoriques.

V.2 📦 CONE : apprentissage pondéré pour la mesure F_β

collaboration autour de deux doctorants

La thèse de Kevin Bascol s'est principalement intéressé à des approches d'adaptation de domaine. Cependant, la tâche portait sur la détection supervisée d'anomalies, et les mesures utilisées pour l'évaluation ont été la précision, le rappel et le score F_1. Nous nous sommes donc intéressés à l'optimisation de la mesure F_β. Étant non-séparable, cette mesure soulève des questions, par exemple en rapport avec l'apprentissage à base de mini-batchs. En collaboration avec un autre doctorant, Guillaume Metzler, et dans le but de combiner les points forts des deux doctorants et de leur donner une occasion structurée de collaborer, nous avons travaillé sur des garanties théoriques et des algorithmes optimisant la mesure F_β en reformulant le problème sous forme d'apprentissage pondéré, où les classes ont des poids différents (et variables). Ces travaux ont donné l'approche présentée dans *From Cost-Sensitive to Tight F-measure Bounds* [21] et surnommée CONE [20].

il existe une pondération optimale

L'intuition de ces travaux est la suivante : en re-ponderant la classe minoritaire (les anomalies) et en faisant un apprentissage pondéré, on peut indirectement optimiser la F_β. Le poids relatif utilisé pour repondérer les classes (anomalies, non-anomalies) est cependant inconnu. En particulier, il ne suffit pas de ré-équilibrer les classes, par exemple en mettant un poids 10 fois plus grand à la classe d'intérêt s'il y a 10 fois moins d'exemples de cette classe.

paramètres t et pondération $a(t)$

Il est en fait possible de démontrer qu'en apprenant suffisamment de modèles avec des pondérations différentes, on peut approcher autant que voulu le modèle qui est optimal en termes de F_β. Ces résultats reposent sur des bornes théoriques qui lient la valeur de la F_β en apprentissage avec la valeur minimale de la F_β que l'on pourrait obtenir (en apprentissage) avec le meilleur modèle de la famille d'hypothèses. Les modèles sont appris avec des pondérations qui dépendent d'un paramètre appelé t. Plus précisément, la pondération est notée $\mathbf{a}(t)$ et définie par $\mathbf{a}(t) = (1 + \beta^2 - t, t)$ où β est le paramètre de la mesure F_β qui nous intéresse. Le coefficient $1 + \beta^2 - t$

est utilisé pour pondérer les erreurs e_1 (faux négatifs) et t pour pondérer les erreurs e_2 (faux positifs). On appelle $\mathbf{e} = (e_1, e_2)$ le profil d'erreur. Avec un t donné, le problème d'apprentissage pondéré considéré minimise donc $\langle \mathbf{a}(t), \mathbf{e} \rangle$. La mesure F_β peut aussi s'écrire en fonction des erreurs (du profil d'erreur), en notant P le nombre de positifs (d'après la vérité terrain) :

$$F_\beta(e) = \frac{(1+\beta^2)(P - e_1)}{(1+\beta^2)P - e_1 + e_2}$$

bornes non informatives

Une démonstration a été apportée par [68] qui prouve qu'en apprenant des modèles pondérés avec une grille suffisamment fine de t, on peut approcher la F_β optimale. Ces travaux présentent deux limites. D'une part, les bornes théoriques sont très larges et, en pratique, avec une grille de 19 modèles comme proposé dans leurs expériences, les bornes sont non-informatives, c'est-à-dire qu'elles nous garantissent uniquement que la F_β optimale est inférieure ou égale à 1 (par définition, sa valeur maximale). D'autre part, du fait de leur imprécision, ces bornes ne permettent pas d'obtenir un algorithme (autre que l'exploration systématique d'une grille de paramètres).

mécanique de demonstration, lien entre F_β et exactitude

Dans ces travaux, nous avons apporté un ensemble d'éléments, à commencer par une interprétation graphique des bornes. Cette interprétation donne son nom, CONE, à la méthode et offre un meilleur support pour suivre les preuves. Un premier point clé de la preuve est relativement intéressant à exposer dans ce document, du fait qu'il pourrait s'appliquer dans d'autres contextes et qu'il est relativement facile à appréhender. C'est le mécanisme qui permet de passer de l'apprentissage pondéré (minimisation de l'exactitude pondérée) à la mesure F_β. Si la F_β était séparable, on pourrait l'écrire en fonction du profil d'erreur, c'est-à-dire dériver mathématiquement des coefficients tels que $F_\beta = c_1 e_1 + c_2 e_2 + c_3$ et optimiser avec les poids (c_1, c_2). La F_β n'est pas séparable… mais est une fraction de formes linéaires (du profil d'erreur). On peut en effet écrire $F_\beta = \frac{c_1 e_1 + c_2 e_2 + c_3}{d_1 e_1 + d_2 e_2 + d_3}$ avec des coefficients donnés. L'astuce est alors d'imaginer une valeur cible pour la mesure F_β, que l'on note t (qui deviendra notre t introduit précédemment) et de se poser la question « peut-on obtenir une valeur de F_β supérieure à t ? ». Cette question est capturée dans l'équation $\frac{c_1 e_1 + c_2 e_2 + c_3}{d_1 e_1 + d_2 e_2 + d_3} > t$ qui après distribution et réorganisation donne $(c_1 - t \cdot d_1)e_1 + (c_2 - t \cdot d_2)e_2 > (c_3 - t \cdot d_3)$. Quand cette dérivation est appliquée avec la formule de la mesure F_β (qui donne des valeurs pour les c_i et les d_i), on obtient alors le vecteur de pondération $\mathbf{a}(t)$ présenté précédemment, et la question devient de savoir si on peut réduire l'erreur pondérée sous un seuil, et justifie donc de minimiser cette erreur. D'autres approches, par exemple basées sur de l'apprentissage par descente de gradient ont aussi utilisé cette formulation, en proposant d'optimiser à la fois le seuil objectif t (et donc $\mathbf{a}(t)$) et l'exactitude pondérée [69].

visualisation de la régularité autours de l'optimal

Les bornes manipulées s'intéressent au lien entre la valeur de F_β obtenue avec un modèle appris par apprentissage pondéré et celle qui pourrait être obtenue avec le modèle donnant le profil d'erreur optimal (au sens de la F_β). Le profil d'erreur optimal est noté $\mathbf{e}^\star$ et est inconnu, mais il serait théoriquement obtenu si on avait la pondération $t^\star$ optimale. On compare ce profil (et sa mesure $F_\beta(\mathbf{e}^\star)$ obtenue théoriquement avec $t^\star$) à celui que l'on a obtenu, $\mathbf{e}$, avec un apprentissage pondéré en utilisant un t particulier. La borne des travaux précédents, [ ] , était essentiellement exprimée à base de quantificateurs et fragmentée en plusieurs morceaux. De plus, les travaux se concentraient sur l'existence du résultat sans s'intéresser à sa finesse. En réinterprétant cette borne et combinant les différents fragments, nous avons proposé de la représenter sur des graphiques F_β en fonction de t, où la borne prend alors une forme de cone. Chaque cone est généré par un point d'observation (une valeur de F_β obtenue par apprentissage pondéré pour un t choisi) et correspond à une région d'impossibilité. Cela matérialise l'intuition derrière la borne : si l'on change un peu la valeur de t, on obtiendra, au mieux, une amélioration en F_β qui est limitée.

des cones plus larges et asymétriques

Notre interprétation sous forme de cone, nous a permis de mettre en évidence à quel point la borne proposée dans [68] n'apportait pas ou très peu d'information dans les cas pratiques utilisés (apprentissage de 19 modèles avec 19 valeurs de t) [ ] . En prenant soin aux détails dans les étapes et en dérivant une borne qui est asymétrique (la pente du cone est différente entre les deux côtés) et qui dépend du profil d'erreur (e_1, e_2), [ ] , nous avons obtenu un résultat beaucoup plus informatif. En gardant l'apprentissage de 19 modèles avec les valeurs de t proposées dans l'article d'origine, nous obtenons avec notre nouvelle borne des garanties très fines, très proches des valeurs empiriques.

exploration guidée par les cones

Ayant dérivé des bornes informatives et calculables, nous avons aussi pu concevoir un algorithme utilisant ces bornes pour guider l'exploration de l'espace des t. L'algorithme procède par une forme de dichotomie : on travaille avec une partition de l'intervalle des t possibles ($[0, t_{max}]$), avec initialement l'intervalle (non-partitionné), et en découpant à chaque fois en deux parties égales le sous-intervalle le plus prometteur d'après la borne. Plus précisément, on teste d'abord la valeur de $t = \frac{t_{max}}{2}$ que l'on utilise pour apprendre un modèle pondéré. On trace alors (virtuellement) le cone correspondant pour déterminer quelle valeur maximale est encore atteignable dans chaque élément (sous-intervalle) de la partition. On sélectionne alors le t suivant comme étant le milieu du sous-intervalle dont la valeur de F_β encore possible est là plus grande (ou d'un des sous-intervalles, par exemple au départ quand la borne n'est pas encore suffisante). Le processus est itéré jusqu'à l'épuisement d'un budget d'apprentissage ou en se basant sur un cri-

tère de faible amélioration de la valeur obtenue ou de la borne. L'algorithme est illustré dans l'article [ ] .

résultats et extensions

Bien que l'exploration systématique d'une grille permette déjà d'obtenir de bonnes performances en termes de F_β (malgré l'absence de garanties des bornes d'origines), notre algorithme améliore la vitesse à laquelle des bonnes valeurs sont trouvées (et donc la qualité de la valeur trouvée en cas de budget limité) et donne des garanties sur l'optimalité de la solution. L'approche proposée peut être généralisée à toutes les mesures exprimables sous formes de fraction de formes linéaires. Dans les pistes théoriques intéressantes, se pose la question de dériver des bornes en généralisation, donc sur un ensemble de test (les bornes présentées ici liant l'exactitude pondérée en apprentissage et la F_β en apprentissage).

V.3 Déséquilibre de classes et plus proches voisins

plus proches voisins

Dans cette section, nous nous intéressons à l'apprentissage à base de plus proches voisins (kNN) dans le contexte de données déséquilibrées. Par simplification, nous considérons une classification binaire même si une bonne partie des observations se généralisent à la classification muticlasses, voir à la régression.

une distribution inconnue par classe

On considére un problème d'apprentissage de classification binaire. Le problème est en fait défini par une distribution pour chaque classe (ou deux distributions de probabilités, avec un poid relatif). Les distributions sont inconnues et le but de l'apprentissage est de généraliser de manière correcte à ces distributions à partir d'un jeu de données d'apprentissage fini. Le jeu de donnée d'apprentissage est un ensemble de points issus de ces deux distributions avec leur étiquette associée.

différentes interactions entre les supports

Si on zoome sur une région de l'espace, plusieurs cas de figure sont possibles. Ces cas sont illustrés dans la figure V.2 qui ne montre que le support de chaque distribution (zone de densité non nulle). Les cas sont les suivants :

A. on est hors du support des deux distributions,
B. on est hors du support d'une distribution, et la décision doit être en faveur de l'autre classe,
C. on est dans une région où les supports se recouvrent,
D. on est à une frontière entre les deux classes, sans que les supports ne se recouvrent.

deux cas d'intérêt (ici)

Dans le cas (A), la décision du classifieur est sans importance dans notre contexte, même si la détection et/ou la généralisation dans ce contexte dit *OutOfDistribution* (OOD) est un domaine de recherche à part entière. Dans le cas (B), on s'attend clairement à ce que l'on puisse apprendre à prédire la classe dont la densité est non-nulle. Dans le cas (C), il y a un recouvrement des densités et l'erreur bayésienne (erreur du classifieur bayésien optimal) est non-nulle. La question est de savoir comment se comporterait un classifieur à base de plus proches voisins dans cette région (avec un nombre limité de points, ou avec un nombre tendant vers l'infini). Dans le cas (D), il y a théoriquement un classifieur parfait, mais la question est de savoir comment un classifieur à base de plus proches voisins approche ce classifieur idéal (selon le nombre de points d'apprentissage).

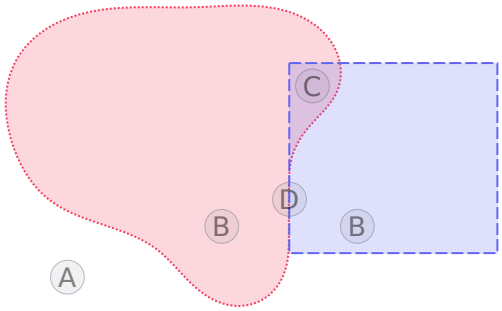


Figure V.2: Support des distributions des deux classes dans un problème de classification binaire. Très localement, on distingue 4 cas. Les cas (A), dit OOD, et les cas (B) qui sont faciles, ne sont pas détaillés. Les cas d'intérêt (C) et (D), avec dans (C) un recouvrement des supports et dans (D) une frontière entre classes, sont l'objet de chapitre.

kNN dans les zones d'incertitude (C)

Nous nous intéressons aux situations (C) et (D), en particulier dans le cas déséquilibré, c'est-à-dire le cas où les distributions ont des masses/valeurs différentes (au moins localement). Dans le cas (C) et avec beaucoup de points, kNN a tendance à respecter les densités relatives entre les points, en particulier avec $k = 1$. Plus k grandit, plus le kNN prédira alors la classe majoritaire et donc plus il s'approchera du taux d'erreur bayésien. Ces propriétés restent vraies quelle que soit la dimension, cependant le nombre de points pour bien remplir l'espace (pour avoir « beaucoup de points ») croit avec le volume et donc exponentiellement avec la dimension.

kNN aux frontières (D)

Dans les situations de type (D), tout kNN (pour $k = 1$ ou autre) se comporte bien asymptotiquement (quand le nombre de points devient grand). On peut par contre observer, dans la figure V.3, que la classe majoritaire est préférée, et ce d'autant plus que le nombre de points est petit. Comme pour (C), ce phénomène s'amplifie avec la dimension de l'espace. Augmenter la valeur de k au delà ne fait qu'amplifier ce phénomène.

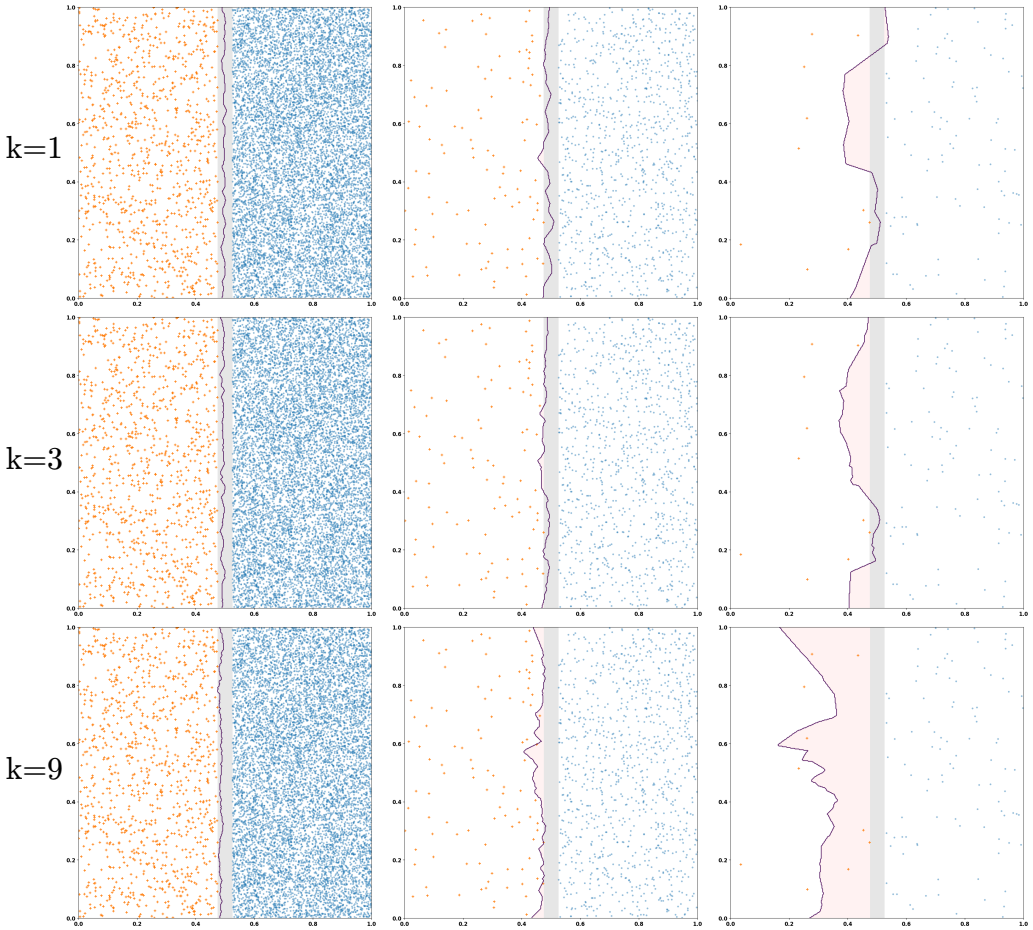


Figure V.3: Évolution de la frontière de décision d'un kNN dans un cas déséquilibré (9 fois plus de points dans une classe) à une frontière entre deux classes séparables, en variant la densité de points et le k de kNN. La zone grise matérialise un espace vide (sans point) à la frontière, ajouté uniquement pour améliorer la lisibilité, en particulier de la figure V.5. Les zones rouges sont les zones d'erreur. L'erreur augmente avec l'augmentation de k et avec la diminution de la densité de points (nombre de points sur l'exponentielle de la dimension).

V.4 Plus proches voisins corrigés pour le déséquilibre

collaboration intra-équipe sur kNN

Ces travaux ont été réalisés en collaboration avec deux doctorants (que je n'encadrait pas) Guillaume Metzler et Rémi Viola, et leurs équipes d'encadrement. La problématique motivant ces travaux était la détection de fraudes supervisée mais où le nombre de fraudes est très petit par rapport au nombre de données normales. Guillaume Metzler avait travaillé sur une approche qui consistait à étendre une ellipse autour de chaque exemple de fraude pour capturer une plus grande région autour de chacune d'elles, sans pour autant augmenter (trop) le nombre de faux positifs (fausse alertes de fraude). Cette collaboration a donné lieu à deux contributions principales. La première, présentée dans cette section, s'intéresse aux approches de type kNN dans le contexte déséquilibré avec [67] étendu dans [63] et présenté à la communauté française dans [66]. La seconde porte sur l'apprentissage de métriques, dans ce même contexte.

générer des points par GAN ?

Notre objectif était de s'inspirer de l'approche à base d'ellipse, mais sans contraindre la forme utilisée pour augmenter la région d'influence des points positifs (fraudes). Notre première idée a été de générer des points dans l'espace entre une donnée positive et les négatifs l'entourant. Pour ce faire, nous avons utilisé une approche de type GAN : en se concentrant sur un point positif et considérant uniquement son voisinage (comme pour l'approche à base d'ellipse), l'idée est d'apprendre un GAN dont le générateur apprend à modéliser la distribution des points négatifs (éventuellement avec un peu de bruit). On peut alors considérer que le discriminateur trace une frontière entre le centre (où il y avait notre point positif mais où il n'y a pas de points négatifs) et les points négatifs autour. Le discriminateur remplace donc l'ellipse, avec une forme beaucoup moins contrainte. Une autre utilisation est de faire un sur-échantillonage, c'est-à-dire augmenter le jeu de donnée avec des points positifs, classés comme *fake* par le discriminateur.

reprendre le problème, analytiquement

L'approche à base de GAN, bien que donnant des premiers résultats intéressants, s'est avérée coûteuse et difficile à stabiliser. Nous sommes donc repartis du problème qui visait à étendre la zone d'influence autour des points positifs (fraudes), mais avec un objectif d'efficacité et de simplicité. L'idée que nous avons proposée est finalement simple : nous utilisons un kNN mais la distance à un point positif est virtuellement réduite.

multiplier la distance aux positifs par γ

On a ainsi un kNN dans lequel la distance entre un point de test x et un point de l'ensemble d'apprentissage x_i dépend de la classe de x_i. Si x_i est un point négatif (normal), on utilise une distance d, par exemple la distance euclidienne, on a alors $d_\gamma(x, x_i) = d(x, x_i)$. Par contre, si x_i est un point

positif (fraude, anomalie), on utilise cette même distance d mais multipliée par un paramètre γ, on a alors $d_\gamma(x, x_i) = \gamma \cdot d(x, x_i)$. La figure V.4 donne une idée de la frontière de décision obtenue avec cette distance d_γ pour différents γ. Le paramètre γ est typiquement inférieur à 1 dans cet exemple, pour réduire la distance perçue. Ce paramètre pourrait être fixé avec un a priori sur le déséquilibre mais, étant dans un contexte supervisé, il est plutôt réglé par validation croisée.

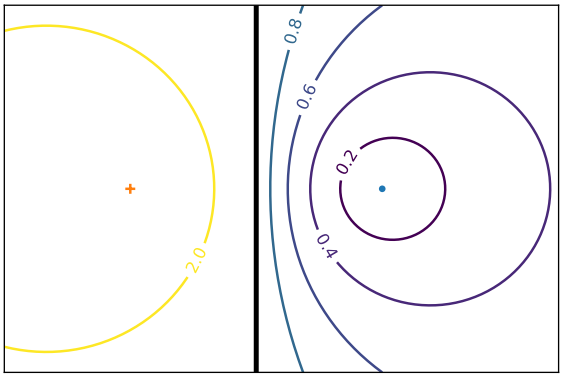


Figure V.4: Frontières de décisions pour différents classifieurs γ-kNN ($k = 1$, γ variable) basé sur la distance d_γ. La valeur sur la frontière de décision est la valeur de γ. Le jeu de données n'est ici fait que de deux points, un positif à gauche et un négatif à droite.

implémentation de γ-kNN

L'implémentation de γ-kNN est très simple et de complexité similaire à kNN (au pire, deux fois plus lent que kNN) et permet de réutiliser des implémentations de kNN (potentiellement rapides et approchées). En effet, pour $k = 1$ et un point requête x, il suffit de chercher le plus proche exemple positif d'une part et le plus proche négatif d'autre part. On fait donc deux recherches de plus proche voisin mais sur des ensembles plus petits. Pour prendre la décision, on compare alors les valeurs, en termes de d_γ, c'est-à-dire en multipliant par γ la valeurs obtenue pour la distance au point positif. Quand k est plus grand, on peut chercher les k plus proches voisins dans chaque ensemble (positif et négatif), multiplier les distances par γ pour l'ensemble positif, puis re-trier ces $2k$ éléments (peu coûteux) pour faire une décision de vote majoritaire à partir des k plus proches voisins.

+rapide

De manière plus optimisée, il est en fait nécessaire de récupérer uniquement le $\frac{k+1}{2}$ème voisin dans chaque ensemble puis de comparer les deux valeurs de d_γ.

effets sur les frontières

La figure V.5 reprend le comportement de kNN à une frontière avec différentes densités de points. Cela illustre que l'utilisation γ-kNN repousse la frontière de décision pour augmenter l'influence des points positifs (avec $\gamma < 1$). Avec un γ petit, γ-kNN tend à encercler les points négatifs : tout

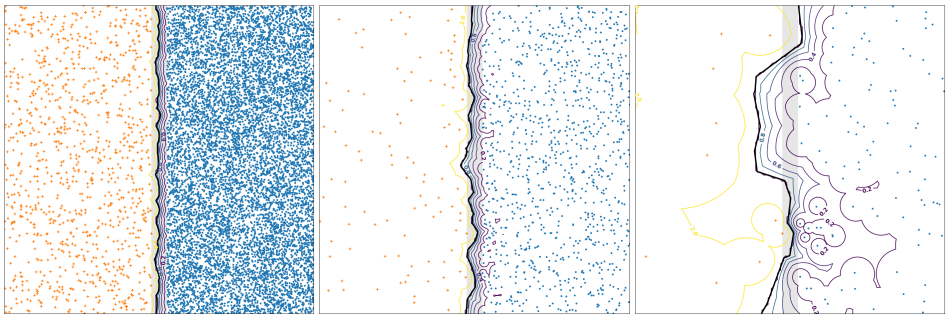


Figure V.5: Évolution de la frontière de décision d'un γ-kNN ($k = 1$) dans un cas déséquilibré (9 fois plus de points dans une classe) à une frontière entre deux classes séparables, en variant la densité de points. La zone grise matérialise un espace vide (sans point) à la frontière, ajouté uniquement pour améliorer la lisibilité. Les différentes lignes de niveaux sont les frontières de décision de γ-kNN pour des valeurs de γ. En noir, kNN ($\gamma = 1$), du bleu au violet $\gamma \in (0.8, 0.6, 0.4, 0.2)$, en jaune une correction dans le « mauvais » sens ($\gamma = 2$), voir figure V.4.

point dans une région éloignée d'un point négatif se retrouve considéré comme positif. L'approche γ-kNN peut donc se comporter en partie comme une détection non-supervisée d'anomalie, pour peu qu'il y ait quelques points positifs à proximité dans le jeu de données.

Une piste pour contrôler le continuum entre classification et détection non-supervisée serait de soustraire un biais pour la distance aux positifs. Cette piste n'a pas été explorée car elle n'apporte pas de garanties, ne serait-ce que sur l'erreur sur l'ensemble d'apprentissage, et le paramètre de biais nécessite d'être exprimé comme une longueur dans l'espace considéré (contrairement à γ qui est un paramètre relatif). L'ajout de ce second paramètre, en plus de γ, semble cependant intéressant à explorer.

extension de γ-kNN

À l'échelle d'un jeu de données, nous avons dérivé des garanties relativement intuitives sur l'amélioration en termes de faux négatifs (pour $\gamma < 1$) ou faux positifs (pour $\gamma > 1$). Les expériences ont aussi montré que γ-kNN améliore les performances en moyenne et qu'il est en plus complémentaire des approches de sur-échantillonage classiques. Nous avons aussi proposé une variante où les points générés par les algorithmes de sur-échantillonage utilisent un γ différent de celui utilisé par les points classiques.

comportement global

Le paramètre γ est typiquement trouvé par validation croisée. Parmi les pistes d'amélioration non explorées, on peut mentionner le fait d'apprendre un champ de γ, c'est-à-dire un γ variable dans l'espace, mais ayant une certaine régularité. Une approche de type processus gaussien (avec *inducing points*) serait probablement adaptée pour éviter un sur-apprentissage local.

choix de γ

V.5 Apprentissage de métrique pour corriger le déséquilibre

restaurer l'anisotropie

L'approche γ-kNN permet de renforcer l'influence d'une classe mais utilise la distance euclidienne comme distance de base. Nous avons donc travaillé à restaurer la capacité à prendre en compte l'importance des dimensions de l'espace d'entrée (comme la méthode d'origine à base d'ellipse). Nous avons alors proposé MLFP (pour *Metric Learning from Few Positive*) introduite dans [65] présenté à la communauté française dans [64].

une métrique pour les positifs

Pour généraliser γ-kNN, nous pouvons le voir comme utilisant une métrique linéaire, avec une matrice égale à $\gamma \cdot I$ pour les positifs (et la distance euclidienne pour les points négatifs). L'idée de MLFP est alors d'apprendre cette métrique M qui sert à comparer un point requête avec un point positif.

apprentissage contrastif

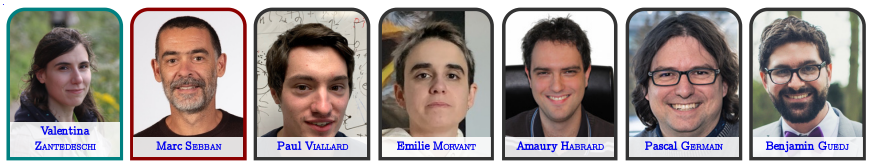
Nous avons formulé un problème d'optimisation en sommant sur deux ensembles de triplets : ceux où 2 points sont positifs, l'autre négatif, et ceux où 2 points sont négatifs, l'autre positif. Le premier point du triplet est considéré comme le point requête et les termes de perte utilisent alors soit la distance euclidienne, soit la métrique (la distance de Mahalanobis), selon la classe de l'autre point (non-requête). Une *hinge loss* avec marge est utilisée pour rapprocher les points de même classe et éloigner les points de classe différente. Nous avons aussi ajouté une régularisation pour inciter M à rester proche de l'identité.

garanties et résultats

À la manière de γ-kNN, nous avons montré que MLFP réduit le taux de faux négatifs dans le cas où les valeurs propres de M sont inférieures à 1. Nous avons dérivé des bornes en stabilité uniforme et sur les taux de faux positifs et faux négatifs pour MLFP. Les expériences sur des taux de déséquilibres variés ont montré que MLFP était meilleur que les autres algorithmes d'apprentissage de métrique, en moyenne. Sur la plupart des jeux de données, soit MLFP soit ImbML [70] (soit γ-kNN) sont les plus performants.

Chapitre VI

Machine learning et garanties théoriques



Publications: [10] [11] [12] [13] [14] [71] [15] [72] [73] [16] [17].

Projets : APRIORI, LIVES, TAUDoS.

---

diversité des types de garanties

En machine learning, il est courant de s'intéresser aux propriétés théoriques d'une formulation ou d'un algorithme. L'éventail des types de garanties théoriques est assez large. On peut par exemple prouver que le minimum d'une fonction relaxée correspond au minimum de la fonction d'origine. On peut aussi prouver qu'avec un ensemble assez grand d'apprentissage on approchera assez bien l'optimal sur des données de test, généralement en supposant qu'elles sont distribuées de la même façon. Les garanties prennent souvent la forme de bornes sur la valeur d'un risque d'intérêt (par ex, risque en test) par rapport à un risque mesurable ou explicitement minimisé.

utilité des bornes ? 🪓 ?

Beaucoup de travaux s'intéressent à prouver des bornes asymptotiques, typiquement quand le jeu de données devient grand. La quantité d'intérêt est alors souvent la vitesse de convergence, c'est-à-dire à quelle vitesse la borne se ressert quand le nombre de points augmente. Quand seul le taux de convergence est considéré (voire même juste l'existence d'une convergence), il est courant de dériver ces bornes de manière grossière (ou « à la hache », comme j'aime le dire). Une grande partie des bornes ne sont donc pas informatives en pratique, par exemple en garantissant que la probabilité d'une mauvaise généralisation est inférieur à 3.14 avec un million de points d'apprentissage

(sachant qu'une probabilité est déjà inférieure à 1).

bornes comme guides

Au delà des bornes asymptotiques, il est possible de dériver des bornes plus fines, plus serrées. Ces bornes sont cependant plus difficiles (et parfois impossibles) à obtenir. Elles nécessitent d'une part de se restreindre à des grandeurs que l'on peut calculer (ou estimer finement), contrairement aux bornes asymptotiques qui peuvent manipuler des grandeurs abstraites (par exemple, le rayon de l'espace, la distance maximale entre deux points, etc). D'autre part, une attention poussée est nécessaire à chaque étape de la dérivation des preuves, en évitant d'ajouter des approximations ou inégalités inutiles. L'effort investi dans ces bornes est bien utile. Ces dernières permettent d'avoir des garanties même sur des jeux de données de taille limitée, d'utiliser leur valeur pour de la sélection de modèle, et de dériver de nouveaux algorithmes soit optimisant directement ces bornes, soit s'en inspirant.

Ce chapitre référence rapidement les autres chapitres de ce document qui contiennent des travaux avec des aspects plus théoriques, puis il se concentre sur deux lignes de travaux : d'une part, la thèse de Valentina Zantedeschi dont un élément structurant était de dériver des garanties dans le contexte de l'apprentissage statistique, et d'autre part des travaux autour de la théorie PAC-bayésiennes en collaboration avec Paul Viallard (dont c'était le cœur de la thèse) et son équipe d'encadrement, ains que les membres du projet ANR APRIORI (porté par Emilie Morvant) dont à nouveau Valentina Zantedeschi après sa thèse.

VI.1 Bornes (informatives) et autres chapitres

accent sur les bornes informatives

Bien que nous ayons aussi dérivé des bornes asymptotiques dans nos travaux, un intérêt particulier a été porté au fait de trouver des bornes informatives. Un cas typique est celui de l'algorithme CONE présenté dans le chapitre V.2. Des travaux avaient montré une convergence vers la F_β optimale quand le nombre de modèles testés augmente, mais n'apportaient aucune garantie dans les cas pratiques considérés. En reprenant l'approche de manière rigoureuse, nous avons produit des bornes beaucoup plus serrées et informatives. À partir de ces bornes, un algorithme de recherche a pu être proposé.

dans les autres chapitres

Une partie des travaux des autres chapitres de ce manuscrit introduisent des garanties théoriques. Certaines sont plus des garanties d'existence ou des garanties asymptotiques (voir IV.2, IV.5.5, V.4, V.5). D'autres prennent la forme de garanties plus serrées et dont la valeur est directement utile ou utilisable pour guider ou proposer un algorithme (voir IV.5.2, V.2).

factor graphs cylciques

Une dernière partie des travaux qui ne sera pas développée ici concerne l'optimisation de modèles de type graphe de facteurs, dans des cas très cycliques. Cela s'applique pour la résolution de problèmes d'optimisation de contraintes distribuées (DCOP) à base de propagation de croyance (*belief propagation*) ou des approches type *MaxSum* pour obtenir un maximum de vraisemblance. Ces travaux ont été réalisés en collaboration, initialement dans le contexte d'un co-encadrement de thèse (le doctorant a dû arrêter en fin de première année pour des raisons personnelles), et ont donné lieu une publication de workshop [71] puis dans une revue [73].

VI.2 Apprentissage local : théorie et algorithmes

thèse de Valentina Zantedeschi

La thèse de Valentina Zantedeschi s'intitulait « *A Unified View of Local Learning: Theory and Algorithms for Enhancing Linear Models* ». Valentina a développé une variété de travaux avec toujours une motivation guidée par la théorie. Un autre élément structurant a été de combiner des modèles locaux. Cela permet d'obtenir des modèles ayant des capacités importantes tant en permettant de dériver des garanties et des algorithmes efficaces.

VI.2.1 Apprentissage de métriques locales

apprentissage local et limitation

L'apprentissage de métriques locales consiste à partitionner l'espace en sous régions et à apprendre une métrique pour chaque région. Si les deux points à comparer (avec la métrique) sont dans la même région alors cela fonctionne bien. C'est par exemple le cas dans les travaux à base d'ellipses locales mentionnés dans le chapitre V.4. Par contre, les métriques locales ne peuvent pas directement servir à comparer deux points qui seraient dans deux régions différentes de la partition [ ] . Nous avons traité ce problème en apprenant la combinaison convexe des métriques locales à appliquer pour chaque paire de régions. Ces travaux ont été publié en conférence [13] et présentés à la communauté française [10]. Un des résultats intermédiaires traitant de la constante de lipschitzité a été extrait sous forme de *preprint* [12].

combinaison convexe de toutes les métriques

La formulation du problème consiste à toujours apprendre une métrique par région mais à ajouter un vecteur de pondération de ces métriques pour chaque paire de régions. Ainsi, pour comparer deux points x_1 et x_2, on regarde les deux régions (potentiellement les mêmes) où ils tombent, disons les régions i et j. La distance/dissimilarité entre les deux points sera une combinaison convexe de toutes les métriques de toutes les K régions, autrement dit, $s(x_1, x_2) = \sum_{z=1}^{K} W_{ijz} s_z(x_1, x_2)$. Si nous avons K régions, nous introduisons donc K^3 nouveaux paramètres (en fait $K^2(K-1)$ puisque l'on force la combinaison à être convexe, c'est-à-dire que les poids somment à

1). L'optimisation se fait alors à la fois sur les K métriques locales et sur le tenseur de pondération W.

double régularisation spatiale

Si aucun a priori n'est mis sur W, rien n'impose au modèle d'apprendre des métriques « locales ». En effet, rien ne dit que la métrique locale numéro z doit être utilisée pour comparer des points de la région z. Nous avons donc ajouté deux régularisations sur W [ ] . La première ajoute une pénalisation ℓ^2 sur les coefficients de W, où chaque dimension est pondérée de façon à inciter la métrique d'une région à ne pas contribuer dans une région éloignée. La notion d'éloignement de région est libre et nous avons utilisé la distance dans un arbre couvrant minimum calculé sur les (centre des) régions. La seconde régularisation ajoute une pénalisation sur W de façon à ce que deux vecteurs $W_{ij.}$ et $W_{i'j'.}$ soient d'autant plus proches (en termes de ℓ^2) que les paires de régions (i, j) et (i', j') sont proches. En pratique, on se retrouve avec des coefficients de similarités $K_{iji'j'}$ pour chaque paire de paires de régions, que nous choisissons de calculer aussi avec la distance dans l'arbre couvrant minimum [ ] . Cette combinaison convexe de métriques par région, en utilisant une mesure de similarité entre régions, pourrait être vue comme une forme discrète d'un processus gaussien. Ceci fait un lien entre ces approches et l'évolution proposée dans le chapitre V.4 pour avoir un γ non constant.

garanties en généralisation

Nous avons dérivé des garanties en généralisation pour la méthode en nous basant sur le cadre théorique de la robustesse algorithmique. Les garanties s'appliquent aussi bien dans le cas de modèles locaux de type ellipse (distance de Mahalanobis) ou de type produit scalaire (formes bilinéaires). Ces travaux ont été appliqués sur de la modélisation de distance perceptuelle entre couleurs et ont permis de réduire l'erreur de prédiction par rapport à une méthode d'apprentissage de métriques locales. Cela montre que l'ajout de (très nombreux) degrés de liberté dans le modèle qui a été fait de manière raisonnée grâce aux régularisations, n'a pas créé de sur-apprentissage et a bien amélioré la qualité prédictive du modèle.

VI.2.2 Apprentissage faiblement supervisé

apprentissage faiblement supervisé et incertitude d'étiquette par point

Dans le contexte de l'apprentissage faiblement supervisé, nous avons introduit la formulation d'un β-risk. L'apprentissage faiblement supervisé englobe par exemple l'apprentissage semi-supervisé (où certains points n'ont pas d'étiquette), l'apprentissage multi-instance (où on donne l'information qu'un point au moins parmi un groupe est d'une certaine classe), l'apprentissage quand l'étiquetage nous dit uniquement quelle proportion de chaque classe il y a dans des groupes du jeu d'entraînement, et plus généralement l'apprentissage quand il y a du bruit dans les étiquettes. Le β-risk est un

substitut (*surrogate*) optimisable qui permet assez directement d'inclure une incertitude sur l'étiquette de chaque point. Associé à ce β-risk, nous avons aussi proposé un algorithme β-SVM permettant de le minimiser [11].

formulation du β-risk

En se basant sur un substitut de fonction de perte F_Φ (par exemple la *hinge*), pour chaque point i, le β-risk considère cette fonction de perte pour chaque étiquette possible et pondère ces valeurs par un vecteur β_i. Les vecteurs β_i sont donnés par la formulation du problème faiblement supervisé. Dans un exemple de classification binaire, si l'étiquette est manquante pour le point i, on a alors $\beta_i = (\frac{1}{2}, \frac{1}{2})$, on combine alors la fonction de perte dans les deux sens. D'une manière générale, la β-hinge est par exemple une fonction en 3 parties linéaires (voir visualisation interactive).

algorithmes et évaluation

Nous avons proposé un algorithme itératif pour l'apprentissage faiblement supervisé, basé sur le β-risk. L'algorithme alterne entre la minimisation du β-risk régularisé et la ré-estimation des valeurs optimales de β. Pour la minimisation du β-risk, nous avons dérivé la formulation duale ainsi que l'algorithme β-SVM. L'approche a été illustrée sur une problématique d'étiquettes bruitées.

VI.2.3 Landmark-SVM et apprentissage multi-vues

approches par landmarks

Les approches à vecteurs de supports, type SVM, sont efficaces dans le cas linéaire mais sont coûteuses quand un noyau est utilisé (et qu'il y a beaucoup de points). Pour réduire ce coût, nous avons proposé une approche à base de *landmarks* (points de repère) qui permet de capturer des non-linéarités tout en contrôlant la complexité. Cette méthode est proche de celles à noyau qui approximent la matrice de Gramm, ou des *inducing points* des processus gaussiens. Une particularité de l'approche est que les *landmarks* utilisées dépendent de la région de l'espace. L'approche de base a été mise à disposition uniquement sur arxiv [14] et une application à l'apprentissage multi-vue a été publié [15].

principe de base de L³-SVM

La formulation proposée s'appelle L³-SVM pour « *Landmark-based Linear Local Support Vector Machines* ». Le principe est d'avoir des modèles locaux pour différentes régions de l'espace, typiquement obtenu par clustering de l'espace d'entrée. Chaque modèle local qui est appris est alors un modèle linéaire sur un espace de caractéristiques qui est un vecteur de similarités entre un point requête et chacune des *landmarks*. Pour créer un problème global, l'ensemble de *landmarks* est commun pour tous les modèles (elles sont typiquement sélectionnées aléatoirement) [ ] .

garanties et expériences

En utilisant le cadre de la stabilité uniforme, nous avons dérivé des bornes en généralisation pour L³-SVM. Par des expériences, nous avons montré que L³-SVM capture bien des non-linéarités et nous avons étudié son comportement face aux hyper-paramètres (nombre de régions, nombre de *landmarks*, noyau). Dans les cas à haute dimension, L³-SVM s'avère plus rapide à apprendre que les autres approches à base de SVM, tout en gardant de bonnes performances en classification.

adaptation multi-vues

Dans le contexte d'apprentissage multi-vues, nous avons adapté L³-SVM en MVL-SVM (multi-vue *landmark* SVM). En apprentissage multi-vues, un point est représenté selon plusieurs vues (ou modalités). Il est possible de comparer des éléments d'une même vue mais la comparaison entre deux vues est généralement impossible. Une première adaptation que nous avons faite est de retirer le côté local de L³-SVM. Nous avons donc un unique modèle. Si l'on imagine un noyau de type RBF (« gaussien », exponentiel carré), l'aspect local a essentiellement comme rôle de potentiellement réduire la complexité, chose qui n'était pas centrale dans ces travaux. La seconde adaptation est d'utiliser un noyau dans chaque vue pour créer un espace de représentation mélangeant des similarités dans différentes vues (puis d'apprendre un séparateur linéaire dans ce nouvel espace) [ ] . Pour des raisons de praticité (et pour la complétion de données manquantes), les *landmarks* sont les mêmes pour toutes les vues.

garanties, expériences, vues manquantes

Nous avons pu dériver des garanties en généralisation pour MVL-SVM, à nouveau en utilisant le cadre de la stabilité uniforme. Nous avons montré que MVL-SVM est robuste aux informations manquantes dans certaines vues. En effet, nous avons proposé une méthode de complétion de données manquantes qui se base sur les informations venant des *landmarks*. Les expériences ont montré que MVL-SVM est plus utile qu'une approximation de Nyström (pour un même rang / nombre de *landmarks*) et bien plus rapide. Notre méthode de complétion de données manquantes basées sur les *landmarks* s'est aussi avérée très efficace, y compris quand elle est utilisée pour ensuite apprendre un SVM simple (même si MVL-SVM reste meilleur qu'un SVM même après complétion de données manquantes).

VI.3 Bornes PAC-bayésiennes

théorie PAC et PAC-bayésienne

La théorie PAC-bayésienne permet de dériver des bornes en généralisation qui sont généralement plus serrées et informatives que celle des cadres théoriques précédents. Le « PAC » veut dire, comme pour les bornes PAC classiques (non bayésiennes), « probablement approximativement correcte ». Le

« probablement » signifie que les bornes tiennent en probabilités, typiquement qu'il existe δ tel que la borne est vraie avec probabilité d'au moins $1-\delta$. Le « approximativement correcte » est le principe même d'une borne, où l'on peut garantir que le résultat tiendra, avec une marge d'erreur bien définie, qui généralement décroît avec la taille du jeu de donnée d'entraînement.

apprentissage d'une distribution de modèle

Le coté « bayésien » vient du fait que ces bornes vont faire apparaître une divergence, typiquement la KL, entre une distribution a priori et une distribution a posteriori. Les distributions en question sont sur l'espace des hypothèses (des classifieurs, des modèles). On s'intéresse donc en particulier aux situations où notre algorithme d'apprentissage ne nous donne pas uniquement un modèle mais plutôt une distribution (a posteriori) sur l'espace des modèles. La distribution a priori a pour seule contrainte de ne pas dépendre des données (y compris d'entraînement).

le PAC-bayésien est-il bayésien ?

Il est important de noter à quel point la théorie PAC-bayésienne se détache de l'inférence bayésienne. Comme mentionné précédemment, la seule contrainte sur la distribution a priori est que l'on ne doit pas utiliser les données d'entraînement pour la déterminer. Elle n'a donc pas particulièrement besoin de représenter une connaissance a priori, elle peut être quelconque et la théorie s'appliquera quand même. De la même façon, la distribution a posteriori, qui est typiquement obtenu par un algorithme d'apprentissage utilisant les données d'entraînement, n'a en aucun cas besoin d'être apprise en suivant la règle de Bayes (en combinant la distribution a priori et les données) pour que la théorie PAC-bayésienne s'applique. En résumé, dans la théorie PAC-bayésienne, les notions d'a posteriori et d'a priori sous-entendent simplement qu'une distribution peut dépendre des données et l'autre non. La théorie est donc applicable dans un contexte d'inférence bayésienne mais est souvent appliquée hors de ce contexte.

a priori informé par les données

Comme les bornes PAC-bayésiennes font apparaître une divergence (typiquement la KL) entre les distributions a priori et a posteriori, il est important d'avoir soit un a priori assez bon, soit un a posteriori qui ne se spécialise pas trop (qui garde une certaine incertitude sur ce qu'est un bon modèle). Une astuce qui tend à rendre les bornes PAC-bayésiennes plus (très) serrées est d'utiliser des données pour se forger un a priori. Ceci n'est pas permis pour appliquer directement la théorie PAC-bayésienne. L'idée est donc de partager l'ensemble d'apprentissage en une partie utilisée pour déterminer un a priori, et une autre partie utilisée pour évaluer les bornes, l'algorithme d'apprentissage pouvant utiliser indifféremment le second ensemble ou l'union des deux. Sans être totalement identiques, ce type de procédure donne des bornes qui ressemblent aux « bornes en validation » qui sont généralement serrées et qui lient la performance sur un ensemble de

test à celle sur l'ensemble de validation (par exemple [74]).

votes de majorités comme distributions

Un cas particulier où le cadre PAC-bayésien s'applique est celui des votes de majorité. Un vote de majorité (pondéré) est en fait une distribution sur l'ensemble des votants. On peut ainsi avoir un vote de majorité a priori, par exemple uniforme, et à partir d'un vote de majorité a posteriori (appris), on peut dériver des bornes PAC-bayésiennes en généralisation. Un type de bornes, dites de premier ordre, peu précises, lient le risque du vote de majorité et le risque en moyenne des votants. Plus précisément cette borne est avec un facteur 2 : le risque du vote de majorité est au pire deux fois plus grand que la moyenne des risques des votants. Un autre type de borne (la C-borne) fait apparaître le désaccord entre votants, c'est-à-dire leur probabilité de prédire des classes différentes, c'est-à-dire leur diversité. En effet, si les votants sont peu divers et prédisent toujours tous à peu près la même chose, le vote de majorité n'apporte rien. Au contraire, si les votants sont plus diversifiés (mais aussi bons en moyenne) le vote de majorité sera meilleur que chaque votant pris individuellement.

VI.3.1 Réseaux de neurones en tant que vote de majorité

Dans ces travaux, nous nous sommes intéressés à dériver des bornes PAC-bayésiennes pour les réseaux de neurones. Le principe est de voir le réseau de neurones comme un vote de majorité. Pour ce faire, nous avons utilisé une décomposition, déjà proposée dans [74], d'un réseau de neurones sous forme de chemins [ ] . Nous avons pu voir cette décomposition comme un vote de majorité et dériver des bornes PAC-bayésiennes. Ces travaux ont été publiés dans un workshop [72].

d'une décomposition en chemin…

La prédiction d'un réseau peut être décomposée sous forme d'une somme sur l'ensemble (exponentiel) des chemins liant une des entrées à la sortie. L'impact d'un chemin ρ dans la somme est un produit de : (1) la valeur de l'entrée qui est à la source du chemin (2) $\vec{w}_\rho$, lui-même produit des poids sur ce chemin, et (3) une activation cumulée le long du chemin qui vaut pour un réseau ReLU soit 0 (si une des activations sur le chemin est nulle) soit 1 (si toutes les activations sont dans le régime linéaire) [ ] .

… vers un vote de majorité

Pour voir cette somme comme un vote de majorité (avec des poids positifs et sommant à 1), on voit chaque chemin comme un votant et on utilise comme poids une version normalisée (à travers tous les chemins) de la valeur absolue de $\vec{w}_\rho$ (la partie indépendante de l'entrée dans la somme). On se retrouve alors avec des votants qui font apparaître un produit avec : une des coordonnées d'entrée, l'activation cumulée sur le chemin (0 ou 1 pour ReLU) (dépendant de toute l'entrée et des paramètres du réseau), le signe de $\vec{w}_\rho$

et la constante de normalisation (pour compenser le fait que l'on prenne la valeur absolue pour avoir des poids positifs et normalisés) [ ] .

a priori et a posteriori

Pour pouvoir appliquer la théorie PAC-bayésienne, l'ensemble de votants doit être fixe pour pouvoir comparer les distributions a priori et a posteriori. Dans la décomposition introduite ci-dessus, les votants dépendent des paramètres w du réseau (à travers la constante de normalisation, le signe de $\vec{w}_\rho$ et l'activation cumulée). L'astuce est donc d'apprendre les poids w (du réseau de neurones) sur un premier sous-ensemble des données. Ces poids servent à définir les votants et la distribution a priori. Sur l'autre sous-ensemble des données on apprend alors des poids v (du réseau de neurones) qui donneront des poids $\vec{v}_\rho$ (du vote de majorité sur les chemins). Comme les votants sont définis avec w, il faut imaginer que le second réseau qui est appris utilise w pour calculer quelles activations sont 0 ou en régime linéaire (pour ReLU), par contre les poids v sont utilisés pour les autres calculs.

bornes PAC-bayésiennes par programmation dynamique

Avec cette formulation, il est alors possible de dériver des garanties en généralisation sur les réseaux de neurones appris avec cette procédure. Bien que l'ensemble des votants (des chemins) soit exponentiel, il est possible d'écrire un algorithme de programmation dynamique pour évaluer efficacement les bornes.

VI.3.2 Vote de majorité stochastique

Dans ces travaux, nous introduisons la notion de vote de majorité stochastique, c'est-à-dire un vote de majorité tiré d'une distribution sur les votes de majorité. Nous montrons que cela permet de dériver des bornes extrêmement précises. Cela vient du fait que l'on arrive alors à optimiser la fonction de perte 0/1 (et non un substitut) en généralisation. Le cadre PAC-bayésien est utilisé pour dériver des bornes à optimiser et des garanties. Ces travaux ont été publiés en conférence [16].

limites des bornes existantes

Les bornes existantes telles que la borne de facteur 2 (premier ordre) ou la C-borne ne sont pas totalement adaptées pour *apprendre* un vote de majorité. En effet, la borne de facteur 2 ignore totalement la diversité et produit donc des algorithmes qui vont simplement sélectionner les votants les meilleurs. La C-borne échoue aussi à sélectionner des ensembles très diversifiés de votants, chose qui est capitale dans le cas de votants plutôt faibles mais complémentaires. Une autre borne dite binomiale approche le vrai risque du vote de majorité (dans le but de l'optimiser efficacement) par des tirages de sous-ensembles de N votants. Plus N est grand, plus l'approximation est bonne pour l'optimisation, cependant, il reste un facteur 2 dans les garanties. Dans l'exemple synthétique classique du jeu de données des deux lunes, et des

souches de décision comme votants (des arbres de profondeur 1, fixés), seule l'optimisation de la borne binomiale permet d'obtenir un classifieur parfait mais les bornes sont 2 à 8 fois plus grandes qu'avec l'approche que nous proposons (qui donne aussi un classifieur parfait).

vote de majorité stochastique

On se place dans le contexte d'un vote de majorité, noté f_θ où θ est typiquement une distribution (un vecteur de poids) sur l'ensemble (discret) des votants. Contrairement à une borne PAC-bayésienne classique qui bornerait un risque en généralisation $R(f_\theta)$, nous nous intéressons à un risque en moyenne sur une distribution $\rho(\theta)$, c'est-à-dire $\mathrm{E}_{\theta\sim\rho}[R(f_\theta)]$. Nous avons ici une modélisation hiérarchique où l'on manipule une distribution sur les distributions, mais qui est stochastique par nature. De façon plus simple (mais qui peut être déconcertante) notre prédiction consiste à tirer un vote de majorité de ρ puis à utiliser ce vote de majorité pour classer un point requête. Si on veut apprendre un ρ, il faut arriver à optimiser cette espérance.

forme close pour le cas conjugué

Nous avons pu transformer analytiquement le risque du vote de majorité stochastique dans le cas où ρ est une distribution de Dirichlet. La loi de Dirichlet est conjuguée au vote de majorité qui est lui-même une distribution sur un ensemble fini, c'est-à-dire une loi catégorique. La loi de Dirichlet a une propriété d'agrégation qui nous permet, pour un point donné, de regrouper les votants en deux groupes, ceux qui votent pour la bonne classe et ceux qui votent pour la mauvaise classe. Cela nous permet d'écrire une forme close pour le risque (en perte 0/1) du vote de majorité stochastique en utilisant la « fonction beta régularisée incomplète ». Cette fonction est compliquée mais calculable et différentiable, permettant ainsi une optimisation par descente de gradient. Il est important d'insister sur le fait que cela donne un moyen d'optimiser directement le risque 0/1 (et non un substitut).

approximation de Monte-Carlo

Pour gérer d'autres cas (autre que Dirichlet) ou pour éviter de manipuler et dériver la fonction béta régularisée, nous avons aussi proposé une approche basée sur un échantillonage. Plutôt que d'intégrer analytiquement $\mathrm{E}_{\theta\sim\rho}[R(f_\theta)]$ on tire T échantillons $\{\theta_t\}_{t=1}^T$ de ρ et on approxime l'espérance par $\frac{1}{T}\sum_{t=1}^T R(f_{\theta_t})$. Pour pouvoir calculer un gradient, il est nécessaire d'appliquer le *reparametrization trick* et de remplacer le risque par un substitut différentiable.

bornes PAC-bayésiennes et évaluation

Nous avons dérivé des bornes PAC-bayésiennes pour le vote de majorité stochastique. Une première borne est pour le cas d'un a priori non-informé (n'utilisant pas de donnée). Une seconde utilise une partie des données pour apprendre un a priori informé par les données. Les expériences montrent la

pertinence de nos deux algorithmes ainsi que la précision des bornes obtenues.

VI.3.3 Bornes désintégrées pour la généralisation

bornes désintégrées

Nous avons exploré un type de bornes en généralisation pour dériver des garanties qui dépendent d'une mesure de complexité arbitraire. Nous avons utilisé le cadre PAC-bayésien, et plus particulièrement les bornes désintégrées. Contrairement aux bornes PAC-bayésiennes qui sont valables en moyenne sur une distribution d'hypothèses, ces bornes PAC-bayésiennes désintégrées sont valables avec grande probabilité sur les hypothèses (donc avec grande probabilité pour une hypothèse aléatoire donnée).

mesure de complexité arbitraire

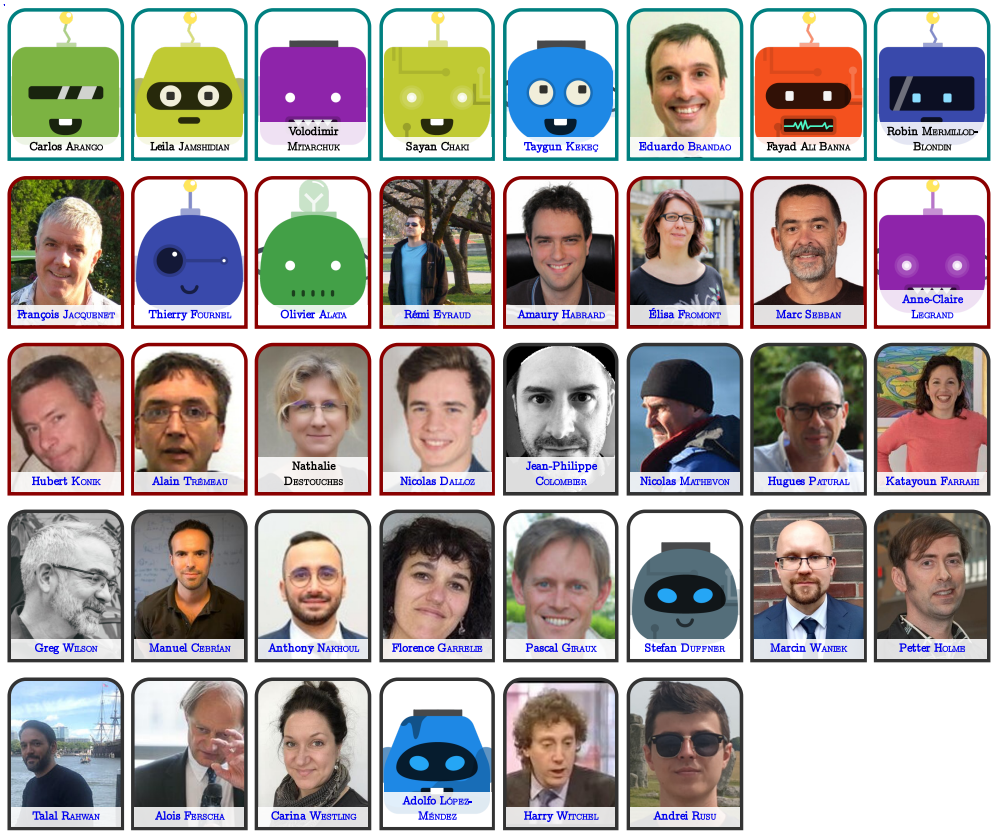
L'originalité de ces travaux est de fournir des bornes qui permettent d'intégrer n'importe quelle mesure de complexité d'hypothèse de notre choix. Une mesure de complexité est une fonction qui essaie de prédire le risque de sur-apprentissage d'une hypothèse. À partir d'une hypothèse et d'un jeu de donnée d'entraînement, une mesure de complexité doit fournir idéalement une estimation de l'écart de généralisation (*generalization gap*). Les bornes proposées, bien qu'assez directes à dériver à partir des travaux existants, sont uniques du fait qu'elles permettent cette intégration de toute mesure de complexité qu'il est possible d'imaginer. Nous avons montré que cette borne permet par exemple d'avoir des garanties en utilisant une mesure de complexité qui est apprise (par méta-learning). Ces travaux ont pour l'instant été présentés (avec succès) uniquement à la communauté française [17] mais ont de bonnes chances de paraître sous peu.

VI.4 Description de longueur minimale (MDL) informée par la tâche

La thèse d'Eduardo Brandao s'est intéressée au domaine du machine learning guidé par la physique et du machine learning pour la physique, en mettant l'accent sur des éléments méthodologiques autour des mesures de complexité et d'entropie. Nous avons en particulier exploré des formulations à base de MDL (*minimum description length*) pour les réseaux de neurones avec le lien entre MDL et entropie conditionnelle. Le principe MDL n'est pas bien adapté pour les modèles sur-paramétrés. Nous avons donc introduit une nouvelle formulation, basée sur des notions de bruit et de signal dont la définition dépend de la tâche. Ces travaux ont été récemment publiés et présentés en conférence [32].

Chapitre VII

Autres travaux



Publications: [75] [76] [77] [78] [79] [80] [6] [7] [81] [8] [82] [9] [83] [84] [31] [33].

Projets : FA40, ROIi, TAUDoS, PIMALEA, BabyCry.

Codes et liens divers

- Plateforme *SoundScapeExplorer* de machine learning et bio-acoustique pour l'analyse des paysages sonores.
  https://github.com/sound-scape-explorer/sound-scape-explorer

- Documentation de la plateforme.
  https://sound-scape-explorer.github.io/

Une conséquence du choix de couvrir exhaustivement les travaux publiés est l'existence inévitable d'un chapitre fourre-tout, que voici. Malgré ce choix, certaines pistes, explorées par exemples sous formes de stages, ou n'ayant pas suffisamment débouché, sont omises de ce chapitre et donc de ce manuscrit. Le but est ici principalement de donner une idée très concise des différents travaux qui sont soit déconnectés des catégories définies par les chapitres précédents, soit difficiles à catégoriser, soit correspondant à des directions de recherches relativement récentes pour moi.

VII.1 📦 Épidémiologie : épidémies et traçage de contact

épidemiologie pré-COVID-19

Lors de mon postdoc, j'ai démarré une collaboration avec Katayoun Farrahi qui faisait son postdoc au même endroit que moi et Manuel Cebrían (à distance) sur le domaine de l'épidémiologie. Cette collaboration était clairement en dehors de nos domaines de recherches du moment. Cette collaboration date de bien avant l'épidémie de COVID-19 et a finalement trouvé une résonance inattendue avec cette crise qui nous a aussi incité à poursuivre un peu plus certains travaux avec de nouveaux collaborateurs.

réseau de contagion et réseau de dépistage

Nous nous sommes initialement intéressés à savoir si un réseau d'interaction créé à partir des échanges de SMS permettait d'approximer les vraies interactions entre personnes (captées par proximité bluetooth) [75]. Dans un contexte de dépistage des cas contact, nous avons ensuite étudié l'impact d'utiliser un réseau imparfait (par exemple issu des interactions mesurées par SMS ou bluetooth) sur l'efficacité d'une politique de dépistage des cas contacts [76],[80].

modèle à compartiments pour le COVID-19

Les modèles épidémiologiques sont généralement des modèles à 3 (ou 4) compartiments. Les individus dans le réseau sont donc dans un des 3 états sain/infecté/rétabli (modèle SIR). Par facilité de calcul, les temps de transitions (temps d'infection ou temps de rétablissement) sont considérés suivre des lois exponentielles. Le cas du COVID-19 est un peu plus compliqué : il nécessite un plus grand nombre de compartiments et aussi de modéliser des durées (par exemple la durée d'incubation) comprises dans un intervalle. Il faut donc manipuler des distributions plus compliquées (par exemple uniforme sur un intervalle) que la distribution exponentielle. Pour correspondre à ces propriétés, via l'encadrement de Andrei Rusu, nous avons proposé un modèle épidémiologique correspondant au COVID-19 [83].

endiguement ou recherche du patient 0

Les derniers travaux partent d'un constat : lors du début d'une épidémie (en particulier), le dépistage des cas contact peut avoir deux objectifs qui ne sont pas forcément alignés : d'une part, essayer de trouver le patient zéro pour mieux comprendre la cause de l'épidémie, et d'autre part contenir la propagation de l'épidémie. Nous nous sommes aussi intéressés, autour du travail de Marcin Waniek, au compromis entre ces deux objectifs [84].

VII.2 Détection de micro-expressions

La thèse de Carlos Arango portait sur la détection de micro-expressions. Ces dernières sont des mouvements subtils et brefs d'une partie du visage. Nous avons développé des approches basées sur la transformée de Riesz appliquée aux images. Cette transformée est une généralisation de la transformée de Hilbert. Nous avons travaillé avec une pyramide de Riesz, c'est-à-dire en travaillant dans l'espace échelle, ou dit autrement avec une représentation multi-échelles. Intuitivement, cette transformation capture les fréquences du signal (de l'image) radialement autour de chaque position. À la manière de la transformée de Fourier qui travaille avec des nombres complexes, la pyramide de Riesz sur des images amène à travailler avec des quaternions. Ces travaux ont été développés pour la détection de micro-expressions d'une part [6],[8], et d'autre part pour leur caractérisation et classification [9]. Ces travaux ont aussi été appliqués dans un contexte clinique, pour la détection de micro-mouvements pour détecter les réactions des patients en sortie de coma [7].

VII.3 Bioacoustique et éthologie

Nous collaborons au sein du l'université avec les bioacousticiens de l'ENES, l'équipe de neuro-éthologie sensorielle du CRNL. Dans cette collaboration, nous nous intéressons aux méthodes de machine learning pour la bioacoustique, que ce soit pour l'exploration de masses de données ou pour la classification/détection automatique. Ces travaux sont structurés autour de financements d'ingénieurs en développement et de postdocs, à l'interface entre bioacoustique et informatique.

structuration autour d'une plateforme logicielle

Il est de plus en plus courant de poser des microphones dans un environnement d'intérêt et de faire des enregistrements (continus ou non) pendant des journées, voir des mois. Les microphones captent donc ce que l'on appelle des paysages sonores, contenant toutes sortes de sons. La première étape vise à donner de la visibilité à cette collaboration tout en structurant les futures contributions. Cette étape se concentre donc sur le développement d'une plateforme logicielle pour l'analyse de ces paysages sonores, disponible sous licence libre. Nous avons donc développé la plateforme SoundScapeExplorer.

extraction, agrégation, réduction, visualisation

La plateforme SoundScapeExplorer repose sur une chaîne de traitement de base qui peut être augmentée avec différentes fonctionnalités (présentes dès maintenant ou disponibles dans le futur). La chaîne de traitement de base reçoit un jeu de données sous forme de nombreux fichiers audio (et l'information de où et quand ils ont été enregistrés) et le traite typiquement comme suit : (1) pour chaque seconde de signal audio, un descripteur est extrait par un réseau de neurones, par exemple appris sur une tâche générique de classification audio, (2) une agrégation (temporelle et potentiellement entre fichiers) de ces descripteurs est réalisée pour avoir une granularité temporelle adaptée aux données à analyser, par exemple pour obtenir 1 point pour chaque 30 secondes de signal, (3) la dimension de ces descripteurs agrégés est réduite à 2 ou 3 pour permettre leur affichage, en utilisant une méthode de réduction de dimension, par exemple UMAP, (4) une interface d'exploration interactive dynamique est générée pour permettre de visualiser le nuage de points en 2D ou 3D, de colorer ce nuage en fonction des méta-données renseignées, ou d'écouter un segment audio correspondent à un point donné. Différentes analyses supplémentaires sont d'ores et déjà disponibles dans la plateforme.

analyse des pleurs des bébés

En collaboration avec Hugues Patural, chef du service réanimation néonatale et pédiatrique du CHU de Saint-Étienne, nous démarrons un projet visant à développer des approches de machine learning pour aider au diagnostique de l'état neurologique des nouveaux nés, nés à terme ou prématurés. Le projet va permettre le recrutement de personnes (entre autres des postdocs) sur tous les domaines nécessaires : captation des données ; collecte, contrôle qualité et pré-traitement de données ; aspects recherche coté bioacoustique et aspects recherche coté machine learning. La plateforme SoundScapeExplorer servira de base et de structure pour les analyses et les nouveaux développements méthodologiques.

VII.4 Intersection physique et machine learning

interaction laser-matière et machine learning

Le laboratoire Hubert Curien est un laboratoire à la fois de physique et d'informatique. Nous avons développé de plus en plus de collaborations à l'interface entre la physique et l'informatique. Les physiciens s'intéressent à l'interaction entre la lumière et la matière. La lumière prend typiquement la forme d'impulsions laser ultra brèves, par exemple des laser femto-secondes, c'est-à-dire dont la durée d'impulsion se mesure en femto-secondes (10^{-15} secondes). Une partie des travaux consiste à comprendre, ou arriver à prédire, l'effet d'un traitement laser sur une surface donnée. En effet, selon la surface et les paramètres du laser, la surface peut se transformer et acquérir des propriétés fonctionnelles intéressantes pour les applications : changement de

couleur selon le point de vue et l'éclairage, hydrophobie, propriétés bactéricides, etc. Les phénomènes physiques mis en jeu sont à des échelles temporelles trop fines pour être mesurées dynamiquement et des échelles spatiales nécessitant des microscopes coûteux en argent et en temps pour l'acquisition des images. Nous travaillons donc dans un contexte où les données sont peu abondantes et où la connaissance physique est imparfaite.

exploiter une connaissance imprécise pour apprendre avec peu de données

Une collaboration bien avancée est celle autour de la thèse d'Eduardo Brandao dont la soutenance a eu lieu en fin 2023. Une partie des travaux d'Eduardo Brandao se sont intéressés à apprendre, à partir de très peu de données expérimentales et d'une connaissance physique approximative (sous la forme d'équations aux dérivées partielles ne modélisant qu'une partie du phénomène physique présupposé). Dans ces travaux, nous avons proposé une approche qui permet d'apprendre en combinant l'a priori physique et le peu de données pour pouvoir prédire le lien entre paramètres du laser et la structure produite sur la surface [31],[33] ⛶ .

Au travers des thèses de Fayad Ali Banna et de Robin Mermillod-Blondin, nous nous intéressons à deux autres problématiques. La première est celle des équations de Maxwell (équations de propagation d'ondes électromagnétiques) avec comme objectif de développer des méthodes de machine learning pour, d'une part, remplacer/accélérer les simulations de propagation de l'onde et, d'autre part, apprendre à prédire directement (sans propagation explicite) l'effet d'une impulsion laser sur une surface en termes d'absorption d'énergie à chaque position sur la surface. La seconde est celle d'arriver à optimiser le choix d'un ensemble de paramètres lasers pour obtenir une palette de couleurs, typiquement dans le contexte de la sécurisation de documents d'identité.

VII.5 Travaux divers au fil du temps

finalisation des travaux en architecture logicielle

Pendant le début de mon postdoc, j'ai finalisé des travaux dans la lignée de ma thèse. Sans détailler l'intégralité des travaux sur ce sujet, le but était de proposer des méthodes de conception de logiciels de façon à ce qu'ils puissent s'adapter dynamiquement à un nouvel environnement. Par exemple, un assistant personnel pourrait s'adapter dynamiquement aux services disponibles dans un bâtiment intelligent, de façon à rendre l'expérience de l'utilisateur la plus simple possible. Les verrous reposaient sur la représentation des éléments logiciels, l'algorithme permettant de raisonner dessus, et la prise en compte de la présence de deux types d'utilisateurs : le programmeur (des différentes briques) du système, et l'utilisateur final novice.

débuts en apprentissage profond

Avant la thèse de Damien Fourure, nous avions recruté un doctorant, Taygun

Kekeç qui a décidé assez vite de changer de sujet de thèse et de laboratoire. Ces travaux ont duré quelques mois, exactement à mon arrivée en poste de maître de conférences. Cela a coïncidé avec nos premiers travaux en apprentissage profond, dans le contexte de la segmentation sémantique d'image (affecter à chaque pixel d'une image, une classe parmi un ensemble fixé). Outre la compréhension de ces approches d'apprentissage profond, nous avons proposé une nouvelle approche permettant d'apprendre d'abord un réseau de segmentation puis d'apprendre un réseau qui apprend à corriger les prédictions du premier et à ajouter de la cohérence spatiale dans celles-ci [77],[78].

influence de la musique sur la marche

Quand nous marchons, nous avons tendance à faire osciller notre centre de gravité d'un côté à l'autre. Nous nous sommes intéressés à mesurer ce phénomène dans le but de mesurer la fréquence des pas des personnes dans des vidéos de basse qualité. L'objectif final était de savoir à quel point la musique dans un espace public influence la fréquence de marche des personnes [79].

sur le développement collaboratif de cours

J'ai été impliqué dans le projet « software carpentry » dont l'objectif principal est d'organiser des ateliers de formations aux outils informatiques pour les scientifiques (non-informaticiens). Dans la continuité de cette implication, nous avons rédigé, avec Greg Wilson et d'autres collaborateurs, un article sur les recommandations à suivre lors du développement de cours de manière collaborative [81].

partie C

Bilan global, perspectives et projets

Chapitre VIII

Recapitulatif et conclusions

objectif manqué

Malgré l'objectif d'exhaustivité, j'ai au final mis de côté la quasi-totalité des projets non-publiants et les stages de masters, et probablement oublié certains éléments importants. Plutôt que de résumer chaque partie de ce manuscrit ou d'essayer de forcer artificiellement mes travaux à rentrer dans une cohérence globale, il me semble plus utile de regarder la trajectoire de mes contributions de manière un peu plus chronologique pour dégager des tendances. Pour différentes raisons, je fais le constat aujourd'hui que ce sont plutôt mes travaux qui ont construit guidé ma direction scientifique, plutôt que l'inverse où mes travaux s'inscriraient dans une grande vision à long terme. Je profite aussi de cette conclusion pour me faire (et partager) quelques réflexions.

VIII.1 Trajectoires scientifiques

d'une thèse en architecture logicielle...

Ma thèse a porté sur un sujet totalement différent de l'apprentissage automatique (architecture logicielle) dans une équipe où certaines personnes faisaient de la perception par ordinateur mais plutôt avec une vision que je qualifierais d'algorithmique. Ayant travaillé sur de l'inférence symbolique, au sens de la logique (à base de règles), ma thèse a fini par me convaincre que tout formalisme de représentation de connaissance se devait d'inclure une modélisation des connaissances incertaines. Je m'intéressais déjà à l'époque, en dehors de mon sujet de thèse, au formalisme bayésien.

...à un postdoc probabiliste

Mon postdoc m'a donné la chance de prendre en main des approches où une modélisation, très souvent formulée en termes de probabilités, donne lieu à un problème d'optimisation. Cette approche résonnait alors beaucoup mieux avec ma vision de la modélisation. Je me souviens avoir été rassuré par celle-ci, évitant ainsi (partiellement) d'avancer à tâtons. Tout le défi devint alors de faire que cette modélisation encode les bonnes

choses et soit adaptée à la fois aux spécificités des données, aux approches d'optimisation/apprentissage, mais aussi aux contraintes de complexité de mise en œuvre. J'ai trouvé dans les modèles probabilistes un formalisme qui m'a suivi jusqu'à aujourd'hui et structure ma façon d'appréhender les problèmes (mais sans être l'unique prisme d'analyse).

deep learning

Mon recrutement en tant que maître de conférences en 2013 a été l'occasion de me mettre au deep learning grâce aux opportunités que l'on m'a offertes. Les différentes collaborations ont été l'occasion de travailler avec tous types de données de capteurs : images, sons, vidéos, séries temporelles. C'est finalement sur les données tabulaires plus « classiques » que j'ai le moins travaillé (ce qui n'est pas très original, le deep learning étant moins utilisé dans ce contexte) mais aussi sur le texte. Les approches de deep learning sont devenues prépondérantes au fil de ces années, et d'une manière générale, l'optimisation stochastique (par descente de gradient) de fonctions fortement non-convexes ne semble pas être amenée à disparaître à court terme.

quantification d'incertitude

Mes intérêts récents autour du deep learning sont principalement de deux types. D'une part, je suis intéressé par les travaux qui rapprochent ces modèles et les formalismes probabilistes. Parmi eux, il y a les développements autour de la quantification de l'incertitude dans le modèle et dans ses prédictions, par exemple la prédiction d'une distribution (e.g., un mélange de gaussiennes) en sortie ou dans une représentation latente d'un réseau.

modularité, nouvel outil de modélisation

D'autre part, il me semble qu'une bonne partie du succès du deep learning vient de la modularité qu'il apporte en permettant de définir et composer des fonctions arbitraires du moment qu'elles sont dérivables. Cela est d'autant plus le cas que cette modularité théorique est supportée par la compétition acharnée entre les frameworks. Cette modularité et la facilité relative avec laquelle il est possible de mettre en œuvre et de tester des nouvelles idées (couches, fonctions de perte, régularisations, tâches prétextes, transfert) est primordiale au succès du deep learning. Je vois surtout cela comme un nouvel outil de modélisation. D'une manière générale, les contraintes architecturales imposent un biais inductif et sont donc un nouveau moyen de réaliser une modélisation : il est possible de spécifier des contraintes assez fortes tout en laissant un fort degré de liberté autour de ces contraintes. Un cas typique sont les auto-encodeurs avec une représentation latente et des parties du réseau très structurées, typiquement les modèles de décomposition de scènes sec. III.7.3 ou les modèles de diffusion.

théorie et apprentissage statistique

En parallèle, j'ai pu monter en compétences dans le domaine de la théorie de l'apprentissage statistique. J'ai eu la chance de me retrouver bien entouré pour la découverte efficace de ces aspects et d'être amené à manipuler rapidement une bonne partie des cadres théoriques de l'apprentissage. Mon arrivée tardive dans le domaine m'a probablement amené un avoir un point de vue pragmatique. Je pense que c'est de là que vient ma volonté de rendre les

développements théoriques aussi concrets que possible. En particulier, je me suis toujours questionné sur l'utilité des garanties théoriques qui ne seraient qu'asymptotiques et non informatives en pratique. Bien que nous ayons dérivé de telles garanties asymptotiques au cours de mes travaux, j'apprécie (et espère avoir contribué à cette orientation) le fait que nous ayons aussi réussi à dériver des garanties informatives y compris dans les cas pratiques, parfois au point que certains reviewers de conférences ont initialement douté des résultats fournis.

adaptation de domaine, transfert

Dès mes débuts en tant que maître de conférences, j'ai commencé à travailler sur des questions d'adaptation de domaine et de transfert. Ce thème a été très structurant et le reste encore dans mon activité, de part sa grande transversalité. En effet, c'est un problème assez large (ou dit autrement, « partiellement bien posé ») autour duquel des développements peuvent être réalisés à la fois sur des méthodes d'apprentissage classiques et avec du deep learning, mais de manière plutôt applicative et de manière très théorique. C'est avec plaisir que je suis retourné sur des formulations plus proches des probabilités (même s'il ne s'agit pas directement d'incertitude) au travers du transport optimal.

apprentissage non-supervisé, fraudes et anomalies

De manière assez durable et sous différentes formes, l'apprentissage non-supervisé (ou faiblement supervisé) a beaucoup motivé mes travaux. Que ce soit avec la modélisation probabiliste qui me suit depuis des années ou la période orientée plutôt sur les données déséquilibrées et la détection de fraudes et d'anomalies, ce domaine s'est montré riche en problématiques intéressantes. Un domaine d'intérêt pour moi concerne les cas où il est nécessaire d'acquérir une compréhension fine du comportement des modèles et des algorithmes, dans l'espace des données ou dans un espace latent. Cette problématique de combinaison d'a priori, de données possiblement biaisées et d'étiquettes potentiellement incomplètes, me semble capitale et au cœur d'une bonne partie de mes travaux.

VIII.2 Encadrements, collaborations, projets et autres réflexions

une variété d'étudiants...

J'ai eu l'occasion d'encadrer et de collaborer avec des doctorants sur une large variété de sujets. J'ai eu également la chance de bénéficier d'une diversité d'expériences d'encadrement. J'ai eu le plaisir d'avoir des doctorants qui ont été des *collaborateurs* à part entière et pour lesquels je pense que l'encadrement consiste essentiellement (au-delà d'être acteur dans la collaboration) à permettre aux personnes de s'épanouir, en les accompagnant dans leur évolution et en les incitant à relever des nouveaux défis. À l'autre

extrémité du spectre, certains doctorants sont plutôt des *étudiants* et ont plus besoin d'un support technique et méthodologique, plus fréquent. La mission est alors plus une mission d'encadrement rapproché pour sensibiliser à l'importance de l'expertise technique et de la rigueur scientifique, et veiller à l'application de ces points.
Dans un cas, on pourrait dire que l'on est sur de la *formation à la recherche* (en tant que domaine) où la personne peaufine son esprit critique et formalise sa démarche scientifique, dans l'autre cas on est plus sur de la *formation par la recherche* (en tant qu'activité) où la personne est essentiellement encore en train de se former mais le fait dans un contexte de recherche (selon le sens de « *à* », « *par* », on pourrait attacher une sémantique très différente à ces expressions).

…un équilibre difficile…

En regardant en arrière, j'ai l'impression d'avoir passé plus de temps avec le second type d'étudiants. Cela a pu être un regret par moments, dans le sens où j'ai été moins disponible pour les autres, en particulier dans certaines périodes où le statut d'enseignant-chercheur peut être très demandant. Avec le recul, il est aussi possible de voir les choses autrement et d'y trouver des aspects positifs. En premier lieu, laisser les doctorants-collaborateurs plus libres (sachant qu'ils en sont capables) les prépare peut-être mieux à la suite de leur carrière. Si c'est le cas, privilégier du temps passé à les encadrer relève plus du bénéfice personnel que de l'intérêt des doctorants. En second lieu, voir un doctorant-étudiant évoluer entre son début de thèse et sa soutenance apporte parfois une impression d'avoir vraiment eu un impact fort.

… dans les co-encadrements

Entre les équipes d'encadrement à deux personnes et les équipes d'encadrement multi-domaines où nous étions plutôt nombreux, la dynamique peut être très différente. Dans un contexte plus restreint, l'encadrement est plus canonique, tel que l'on peut l'imaginer, par exemples dans les paragraphes ci-dessus. Dans un contexte avec beaucoup d'encadrants, chaque personne apporte un point de vue, une expertise, mais en tant que groupe nous avons aussi collectivement la mission de faire que le doctorant arrive à devenir expert et moteur sur son sujet (malgré la quantité et la diversité d'expertises qui l'entoure). Chaque encadrement est au final une unique combinaison entre l'étudiant, l'équipe d'encadrement, et le contexte lié au financement et au sujet. Je pense avoir eu un certain aperçu de cet espace combinatoire, sans pour autant être très confiant sur la capacité de généralisation que ces expériences m'apportent.

dépôts de projet et collaborations

Le dépôt de demande de financement et de projet et le fait de les mener à bien fait partie du quotidien. Je suis très content des projets et collaborations auxquels j'ai pris part, et ce sont ces éléments qui m'ont amené ici. Cependant, j'ai parfois eu l'impression de manquer des collaborations qui

me motivaient vraiment, et ce pour cause de rejet de projet. Cela a pu, par moments, me laisser une impression d'un manque d'auto-détermination. Si l'on combine l'impact négatif d'un manque d'auto-détermination, le fait que le métier est très exigeant, le faible taux de financement des projets et parfois les difficultés de recrutements (en particulier en postdoc), la question de l'efficacité de ces outils de financement se pose.

D'un point de vue personnel, la question se pose sur l'implication dans les dépôts de projets et dans les projets eux-mêmes. J'ai été et suis toujours impliqué dans de nombreux projets, sur des sujets assez divers, et je trouve cela très stimulant, mais la question est de savoir si mes contributions n'auraient pas été plus importantes avec une implication dans un plus petit nombre de projets. Hélas, cette situation n'est pas facilement contrôlable. Les taux d'acceptation très faibles rendent impossible toute prédiction sur la charge (en projets) à un ou deux ans, et il y a clairement une incitation à se tromper du côté du « trop » : une absence de projet entraîne une perte de dynamique alors que la surcharge permet toujours de travailler mais en mode légèrement dégradé et au prix de concessions sur d'autres aspects.

moins mais mieux ?

J'exclus du raisonnement suivant les possibles sujets majeurs et très spécifiques. Si le but de réduire les impacts négatifs associés aux dépôts infructueux de projets (temps passé, etc), il serait probablement intéressant de mettre en application des principes de sélection aléatoire de lauréats. Dans les expériences de ce type, cette selection est réalisée après une unique évaluation visant uniquement à assurer une qualité scientifique suffisante, sans essayer d'interclasser toutes les propositions. A priori, ces principes aléatoires financent une recherche plus diversifiée, moins sensible aux effets de modes, et évitent une mise en concurrence qui ne fait qu'augmenter le temps passé (et donc « perdu » à 76%, si l'on se base sur les résultats ANR 2022).

financement par sortition ?

Chapitre IX

Perspectives

une multitude de pistes

Il m'est extrêmement difficile de lister toutes les perspectives possibles. Le nombre de post-its dans mon bureau avec des idées (et le nombre d'emails à moi-même) n'a fait que grandir au fil des ans. Même si une bonne partie des idées ne surviv(rai)ent pas à la recherche bibliographique ou se retrouve traitée par la communauté dans les années (mois, semaines, jours) suivantes, la période d'écriture de ce manuscrit a certainement contribué à augmenter le nombre de perspectives potentielles.

projet IUF ∪ reste

Une première partie résume ici mon projet d'IUF, dont l'écriture date maintenant d'un an et dont la forme est assez contrainte. Les autres parties se libèrent un peu plus de ces contraintes. L'exercice d'écriture de dossier IUF a finalement rendu ce projet structuré, dense et concret. D'un certain coté, j'irai même jusque dire que ce projet est trop structuré (pour être complètement suivi). Pourquoi ? Une présentation inspirante, donnée à la cérémonie d'installation de l'IUF, nous invitait à prendre la mesure de la responsabilité qui vient avec cette reconnaissance. Cela a vraiment résonné avec mon état d'esprit : au-delà d'exécuter un projet écrit plusieurs années à l'avance, voire à la place de l'exécuter, le temps libéré par l'IUF se doit d'être un temps bien utilisé pour renforcer des aspects qui sont indépendants du projet. Entre autres, en tant que note à moi-même, voici les points qu'il me semble important de cultiver : la créativité, la consolidation de l'existant, l'exigence dans la méthode scientifique et l'évaluation, l'ouverture (partage, collaboration), la reconnaissance de toutes les contributions, et la multidisciplinarité nécessaire pour faire face aux énormes défis du monde actuel.

IX.1 Généralisations du transport optimal (IUF)

Le projet se construit principalement sur les travaux autour de la thèse de Tanguy Kerdoncuff dont les travaux sur OTT (Optimal Tensor Transport, voir sec. IV.5.6). Il est articulé autour de 4 directions de recherche (notées

Dir1 à Dir4). Ces 4 directions sont brièvement exposées ici, la version intégrale étant disponible en ligne.

IX.1.1 Dir1 : passage à l'échelle des extensions d'OTT

Bien que la formulation OTT soit très générique, elle peut être étendue pour inclure encore plus de principes dans un cadre général unifié. On peut citer le cadre multi-marginal, la relaxation marginale [85], l'inclusion de l'information de classe, l'inclusion du déséquilibre de classe [86] ou l'utilisation de l'apprentissage de métriques [28]. **Proposer un cadre généralisé intégré** peut transformer le transport optimal en une nouvelle façon de modéliser (comme les modèles graphiques pour les modèles probabilistes), en particulier pour l'apprentissage par transfert, et peut servir de représentation pivot pour de nouvelles contributions orthogonales telles que de nouveaux types de plans de transport (par exemple, pour les séries temporelles), des types de contraintes (par exemple, les marginales relaxées), des algorithmes pour résoudre le problème.

L'un des principaux défis de l'OT, et surtout de ses généralisations comme GW et OTT, est le passage à l'échelle. Les approches rapides existantes peuvent être classées en trois catégories : les approches « sliced » basées sur des projections aléatoires [62] [87] [29], les approches stochastiques [29] [88] basées sur l'échantillonnage, et les approches hiérarchiques [89] [90]. **La conception d'approches stochastiques hiérarchiques avec quantification de l'incertitude**, combinée à la complexité en temps de $N.\log(N)$ par projection, peut réduire la complexité pratique des problèmes d'OT généralisés (approximatifs). Intuitivement, une approche hiérarchique permet de traiter d'énormes ensembles de données en les simplifiant à une taille intermédiaire. Sous cette forme simplifiée, les algorithmes actuels (comme EGW) ont toujours une complexité élevée au-dessus de N^3 pour le cas GW le plus simple et l'échantillonnage est donc le moyen de réduire cette complexité au prix d'une plus grande incertitude. En modélisant mieux cette incertitude, par exemple dans une formulation bayésienne [91], nous pouvons espérer obtenir des algorithmes rapides plus robustes. Les projections aléatoires peuvent jouer un rôle à différents niveaux dans un cadre généralisé, pour réduire davantage la complexité : il peut être utilisé comme une étape interne comme dans PoGroW [29], mais nécessite un schéma de sélection (par exemple, découpage maximal) ou de vote (pour agréger correctement les différentes projections), ou il peut être utilisé comme une méthode d'initialisation aléatoire pour une approche qui l'affinerait davantage (par exemple, pour explorer localement le polytope des plans de transport admissibles).

Une autre direction pour améliorer le passage à l'échelle des problèmes tels qu'OTT est celle proposée dans DifFused Gromov Wasserstein (DFGW) [92]. Introduit dans un cadre de GW fusionné, DFGW exploite les informations des arrêtes (information du graphe induit par les coûts GW) pour diffuser les

caractéristiques des nœuds (information OT) et résout ensuite un problème OT traditionnel, qui est beaucoup moins complexe que GW. **La généralisation de DifFused-GW à OTT** ouvre de nombreuses questions sur la manière de généraliser la diffusion des graphes à des tenseurs d'ordres supérieurs.

IX.1.2 Dir2 : garanties de généralisation et OT

En tant que distance entre des distributions, à la fois continues et empiriques, la distance de Wasserstein a été utilisée dans la plupart des situations où la divergence de Kullback-Leibler (KL) apparaît habituellement. En effet, la divergence de KL est omniprésente dans l'apprentissage statistique mais peut être indéfinie (par exemple, une valeur infinie, selon les supports des distributions). La distance de Wasserstein (au prix d'une notion de géométrie sur l'espace sous-jacent) donne des valeurs pertinentes même pour des supports disjoints et fournit donc une meilleure information sur le « gradient » pour les algorithmes d'optimisation. En tant que distance robuste, la distance de Wasserstein a été utilisée dans de nombreux domaines : adaptation de domaine, réseaux adversaires génératifs, auto-encodeurs (WAE), inférence variationnelle, etc. Il existe des travaux théoriques qui dérivent des garanties de généralisation basées sur la distance de Wasserstein (par exemple, nos travaux sur l'apprentissage de métrique [28]).

Lors de la conception de nouvelles formulations et de nouveaux algorithmes d'OT, dans des approches comme OTT (par exemple, qui utilisent une descente stochastique de Frank-Wolfe ou une descente en miroir projetée), nous parvenons à prouver la convergence de l'algorithme, mais aucun aspect de généralisation n'est pris en compte. Le problème de transport optimal standard (et entropique) a été étudié sous cet angle (par exemple, [93] [94]). Cependant, **dériver des bornes de complexité d'échantillon et de généralisation pour le transport structuré (GW, OTT)** est un problème ouvert et difficile.

La formulation du problème du transport optimal d'une manière probabiliste ouvre le raisonnement sur l'incertitude et la généralisation. En particulier, une formulation bayésienne peut être facilement réutilisée de [91] (qui l'utilise pour gérer des fonctions de coûts stochastiques). En effet, le transport optimal entropique avec distributions empiriques implique naturellement des distributions catégorielles (multinomiales) et la formulation bayésienne y ajoute des a priori de Dirichlet. En général, **l'exploitation du cadre PAC-Bayes (PB) pour le transport optimal** est une direction très prometteuse. La méthode de Sliced-Wasserstein, qui moyenne des projections aléatoires, se prête assez directement à la vision de vote PAC-bayésien [95]. Le cadre PB n'est pas limité à ce contexte de vote et il est capable de produire des garanties en généralisation extrêmement fines dans le cas de la paire (conjugée) catégorique/dirichlet, comme nous l'avons montré dans [16], et il peut très probablement être adapté à la formulation bayésienne de l'OT.

IX.1.3 Dir3 : OT pour les représentations latentes structurées

Cette direction de travail est motivée par un domaine d'application particulier mais peut suggérer des développements pour Dir1 et Dir2. Cette direction est liée à la thèse de Sayan Chaki dans laquelle des auto-encodeurs avec des espaces latents très structurés sont utilisés [96] [97]. L'espace latent représente les propriétés des objets, l'encodeur est un détecteur d'objets et le décodeur un moteur de rendu (stochastique et differentiable). L'idée derrière ces modèles est de décomposer automatiquement un ensemble d'images en leurs objets récurrents constitutifs. La question se pose de comment **l'utilisation de l'OT pour régulariser et transférer des représentations latentes structurées** peut accélérer et améliorer l'apprentissage non supervisé et auto-supervisé.

IX.1.4 Dir4 : OT, EDP et processus de diffusion

Cette direction de travail consiste à **explorer et approfondir le lien entre OT et les équations aux dérivées partielles (EDP)**, ce qui est une tâche stimulante impliquant beaucoup de revues de littérature, d'analyse et de synthèse. Ayant commencé à travailler, au sein du laboratoire, avec des physiciens travaillant sur la théorie et les applications de l'interaction laser-matière, les EDP sont omniprésentes : elles peuvent être connues, elles peuvent devoir être apprises à partir de données ou leurs formes générales peut être supposées et utilisées comme guides/régularisations pour l'apprentissage. En tant que tel, le rapprochement des EDP et de l'OT pourrait être un moyen d'unifier les connaissances physiques (connaissance additionnelle du point de vue de l'apprentissage automatique) avec les données utilisées lors de l'optimisation pour l'apprentissage automatique.

Diverses recherches sur le transport ont déjà souligné plusieurs liens avec les équations différentielles (par exemple, [98] [99] [100]) mais la relation avec l'OT structuré (GW, OTT) reste à étudier. De plus, récemment, des méthodes très efficaces de modélisation de la densité (par exemple pour la génération d'images) ont utilisé des modèles de diffusion (sous forme d'EDP stochastiques ou déterministes) pour générer des échantillons d'entraînement afin d'apprendre le processus de diffusion inverse (débruitage). À ce titre, une direction de recherche prometteuse consiste à **expliciter le lien entre les modèles d'OT et de diffusion**. Les approches de diffusion optimisent en fait le même objectif que les modèles génératifs basés sur le score et plus précisément les approches de *denoising score matching*. Il y a un parallèle à établir entre le réseau de score qui est appris dans ces méthodes et le (gradient) de la fonction de la dualité de Kantorovich-Rubinstein qui donne par exemple le réseau « critique » utilisé dans les Wasserstein-GAN. Une quasi-équivalence existe [101] (prouvée mathématiquement pour un cas particulier et vérifiée empiriquement dans tous les cas testés) entre les modèles de diffusion et le transport optimal vers une distribution normale unitaire.

Ce lien incite à se demander si le problème général de transport optimal (et d'autres problèmes structurés de transport optimal) entre deux distributions est également équivalent à deux processus de diffusion (indépendants) et, si ce n'est pas le cas, si des limites peuvent être dérivées et utilisées pour l'initialisation ou comme étape élémentaire dans un algorithme itératif. Les liens avec les récentes limites PAC-Bayes basées sur les trajectoires sont également prometteurs à explorer [102] en lien avec Dir2.

IX.2 Sciences (physique) et machine learning

interaction lase-matière et machine learning

Avec la création au laboratoire d'une équipe-projet Inria à l'interface entre apprentissage automatique et interaction laser-matière, une partie grandissante de mon activité va naturellement s'orienter dans cette direction. Cette thématique de recherche implique la manipulation d'équations aux dérivées partielles (EDP). Des liens assez directs apparaissent avec mon projet sur le transport optimal et d'une manière générale au carrefour entre apprentissage automatique, transport optimal, EDP et les modèles de diffusion. D'une manière encore plus générale, cela sera l'occasion de développer des approches à la frontière entre l'apprentissage automatique, la géométrie des données (et des dynamiques d'apprentissage) et la physique.

transfert avec peu de données, guidé par la physique

Les premières directions sont dans la continuité de ce qui a été ou est actuellement fait dans les thèses d'Eduardo Brandao et Robin Mermillod-Blondin où le but est essentiellement d'utiliser l'apprentissage automatique pour apprendre à prédire l'impact des paramètres d'un laser sur les propriétés d'une surface (couleur, forme à l'échelle nanométrique, absorption d'énergie, etc.). Comme la quantité de données est faible, un accent particulier est mis sur l'adaptation d'un modèle à de nouvelles situations (nouveau matériau, nouveaux paramètres laser, etc.). C'est une des pistes sur laquelle se concentrera le postdoc d'Eduardo Brandao.

connaissance scientifique assistée par machine learning

Dans ce contexte physique/machine learning, une direction que je compte développer plus, et qui est un second objectif de la thèse de Fayad Ali Banna, est la découverte de connaissances scientifiques assistée par apprentissage automatique. Cette problématique très intéressante pose des questions multiples. Quelle connaissance experte peut ou doit être intégrée, et sous quelle forme ? Quelle représentation doit utiliser l'algorithme pour produire une connaissance utile pour la science (prédictive, compréhensible, vérifiable/testable) ?

effet de doubles impulsions laser

Pour donner un exemple concret, prenons celui de la nano-structuration

d'une surface par laser qui se fait grâce à une succession de *doubles impulsions* laser. La propagation du laser (onde-éléctromagnétique) et son interaction avec la surface sont très bien décrites par les équations de Maxwell. Ces équations permettent de simuler comment le matériau absorbe de l'énergie en fonction de sa forme et de l'impulsion laser (sur des temps physiques très brefs). Sous l'effet de cette énergie, le matériau subit des transformations et entre partiellement en fusion et subit des effets de convection. Ces phénomènes se produisent pendant un temps (court mais moins brefs, de l'ordre d'une dizaine de picosecondes) qui s'écoule avant que la seconde impulsion laser n'arrive sur la surface. Même s'il est possible d'imaginer une fusion du matériau, le comportement précis du matériau n'est pas bien connu dans cette période, et il n'est observable. Il est certain que ce comportement est primordial à la structuration de la surface puisque l'effet final observé (après une double impulsion) dépend du délai entre les deux impulsions.

découverte du processus d'auto-organisation

L'objectif des travaux est d'intégrer un simulateur des équations de Maxwell pour modéliser l'absorption d'une surface donnée. À partir de la forme de la surface et de son absorption, il faut alors apprendre un modèle qui prédit l'évolution de la surface dans le temps, ce modèle devant respecter des contraintes physiques mais aussi prendre une forme accessible aux experts du domaine : les EDP sont typiquement un formalisme pivot intéressant dans ce contexte. L'apprentissage du processus d'évolution (inconnu) doit se faire grâce à la supervision disponible (les observations), c'est-à-dire uniquement l'effet combiné du processus impulsion-absoption-évolution-impulsion-absoption-évolution. Cet apprentissage requiert de « raisonner » sur une absorption (équations de Maxwell) qui a lieu sur une surface qui a déjà évolué (et donc qui est inconnue). C'est pourquoi l'apprentissage doit simuler le processus d'absorption en tant que sous-routine, qui en tant que telle est trop coûteuse en temps de calcul.

apprentissage d'un substitut efficace

C'est en partie dans ce but que nous avons commencé par l'apprentissage d'un substitut, différentiable et beaucoup plus efficace, des équations de Maxwell et/ou directement du processus d'absorption. Ce substitut servira aussi aux physiciens pour explorer dynamiquement l'impact de la topographie d'une surface, des paramètres laser, du matériau, dans le but de gagner en compréhension des phénomènes et de proposer de nouvelles expériences.

IX.3 Pluridisciplinarité et bioacoustique

pluri-disciplinarité et machine learning

Les collaborations en bio-acoustique sont fertiles en termes de financements (et recrutements) et d'interactions avec des personnes passionnées et ayant un esprit scientifique développé. Dans le monde de la recherche, la pluridisciplinarité est souvent encouragée, à juste titre, du fait de l'impact potentiel sur les domaines de recherche. Bien que ces collaborations soient enrichis-

santes et stimulantes, il n'est cependant pas toujours évident de réaliser des contributions fondamentales en apprentissage automatique. Dans les collaborations pluridisciplinaires, l'apprentissage automatique est généralement très vite « utile » pour l'autre domaine. Le transfert dans l'autre sens arrive souvent un peu plus tard : c'est par la compréhension plus fine des problématiques que de nouveaux verrous en apprentissage automatique peuvent apparaître, et donc stimuler le développement de nouvelles approches et méthodes.

des verrous à découvrir

La bioaucoustique est un domaine riche en données et en problématiques, que ce soit autour de la représentation de données, de la segmentation, de la classification ou de la visualisation. J'ai eu plusieurs fois l'occasion de voir que faire des aller-retours entre applications et travail plus fondamental est une bonne source de motivation et de nouveaux développements : en poursuivant une application, en mode « ingénierie » (en appliquant les méthodes existantes qui fonctionnent), un nouveau verrou scientifique apparaît et pour le lever il faut se plonger dans les détails des dernières avancées scientifiques et proposer de nouvelles choses qui peuvent parfois devenir très techniques. Il est difficile, et c'est bien l'intérêt de l'exercice, d'anticiper les verrous qui seront identifiés par les collaborations en bioacoustique ou autour de l'ingénierie de surface. C'est uniquement en ayant confiance dans ce processus pluridisciplinaire que l'on peut espérer arriver à identifier ces verrous dans le temps.

apprentissage actif

Une première direction qui revient souvent à travers les applications est l'apprentissage incrémental et actif. Cet apprentissage actif est encore utile dans un contexte classique où l'on veut annoter le moins d'exemples possibles pour apprendre par exemple un classifieur ou un détecteur de son. Ce côté actif est aussi de plus en plus nécessaire et demandé dans un contexte de transfert d'un modèle vers une nouvelle situation, que ce soit du transfert classique ou l'adaptation de modèles de fondation. Le domaine de l'apprentissage actif pourrait donner l'impression que tout y a été fait, mais c'est un domaine très actif… L'optimisation bayésienne, typiquement utilisée pour l'optimisation d'hyper-paramètres ou pour les plans d'expériences, partage aussi des éléments avec l'apprentissage actif, malgré des différences dans la formulation. Je suis convaincu qu'il y a des avancées à réaliser sur le domaine de l'apprentissage actif et de l'optimisation bayésienne, en particulier en mêlant les spécificités des algorithmes utilisés (souvent une descente gradient stochastique), le cas du transfert, les garanties en généralisations et la quantification d'incertitude.

réduction de dimension et transport optimal

Une autre direction, directement issue des problématiques liées à la bioacoustique, concerne la visualisation de données, au sens de la réduction de

dimension. L'approche UMAP (*uniform manifold approximation and projection*) est aujourd'hui beaucoup utilisée, y compris dans la plate-forme que nous avons en développement. Certaines de ces approches de réduction de dimension, mais aussi les approches de clustering, peuvent être formulées comme un problème de barycentre de Wasserstein. Cela fait le lien avec mes perspectives autour du transport optimal. Les généralisations du transport (GW, Fused-GW, DiffusedGW, OTT, …) seraient de bons candidats pour formuler des méthodes de réduction de dimension (voire avec du clustering simultané) prenant en compte plus d'informations ou de contraintes, et permettant d'être plus rapides et plus stables.

The end.

Ce paragraphe conclut le chapitre de perspectives et donc ce manuscript, en espérant que la lecture de ce dernier vous ait été au moins en partie utile et intéressante.

partie D

Bibliographie

[1] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Trémeau, et C. Wolf, « Mixed Pooling Neural Networks for Color Constancy », in *IEEE International Conference on Image Processing*, 2016.

[2] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, et C. Wolf, « Segmentation de Scènes Extérieures à Partir d'ensembles d'étiquettes à Granularité et Sémantique Variables », in *RFIA 2016*, 2016.

[3] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Trémeau, et C. Wolf, « Semantic Segmentation via Multi-task, Multi-domain Learning », in *S+ SSPR 2016 The Joint IAPR International Workshops on Structural and Syntactic Pattern Recognition (SSPR 2016) and Statistical Techniques in Pattern Recognition (SPR 2016)*, 2016.

[4] D. Fourure, R. Emonet, E. Fromont, D. Muselet, N. Neverova, A. Trémeau, et C. Wolf, « Multi-Task, Multi-domain Learning: Application to Semantic Segmentation and Pose Regression », *Neurocomputing*, 2017.

[5] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Trémeau, et C. Wolf, « Residual Conv-Deconv Grid Network for Semantic Segmentation », in *Proceedings of the British Machine Vision Conference, 2017*, 2017.

[6] C. Arango, O. Alata, R. Emonet, A.-C. Legrand, et H. Konik, « Subtle Motion Analysis and Spotting Using the Riesz Pyramid », in *VISIGRAPP 2018*, 2018.

[7] A. Bertholon, C. Duque, O. Alata, R. Emonet, A. Legrand, H. Konik, et P. Giraux, « Validation in Healthy Subjects of a Clinical Protocol for the Evaluation of Facial Micro-Expressions in Severely Brain Injured Patients Awakening from Coma », *Annals of Physical and Rehabilitation Medicine*, 2018,

[8] C. Duque, O. Alata, R. Emonet, A.-C. Legrand, et H. Konik, « Micro-Expression Spotting Using the Riesz Pyramid », in *WACV 2018*, 2018.

[9] C. A. Duque, O. Alata, R. Emonet, H. Konik, et A.-C. Legrand, « Mean Oriented Riesz Features for Micro Expression Classification », *Pattern Recognit. Lett.*, 2020, https://doi.org/10.1016/j.patrec.2020.05.008

[10] V. Zantedeschi, R. Emonet, et M. Sebban, « Apprentissage de Combinaisons Convexes de Métriques Locales Avec Garanties de Généralisation », in *CAp2016*, 2016.

[11] V. Zantedeschi, R. Emonet, et M. Sebban, « Beta-Risk: A New Surrogate Risk for Learning from Weakly Labeled Data », in *NIPS 2016*, 2016.

[12] V. Zantedeschi, R. Emonet, et M. Sebban, « Lipschitz Continuity of Mahalanobis Distances and Bilinear Forms », *arXiv preprint arXiv:1604.01376*, 2016, https://arxiv.org/abs/1604.01376

[13] V. Zantedeschi, R. Emonet, et M. Sebban, « Metric Learning as Convex Combinations of Local Models with Generalization Guarantees », in *CVPR2016*, 2016.

[14] V. Zantedeschi, R. Emonet, et M. Sebban, « L3-SVMs: Landmarks-based Linear Local Support Vectors Machines », *CoRR*, 2017, http://arxiv.org/abs/1703.00284

[15] V. Zantedeschi, R. Emonet, et M. Sebban, « Fast and Provably Effective Multi-view Classification with Landmark-based SVM », in *ECML-PKDD 2018*, 2018.

[16] V. Zantedeschi, P. Viallard, E. Morvant, R. Emonet, A. Habrard, P. Germain, et B. Guedj, « Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound », *NeurIPS*, 2021.

[17] P. Viallard, R. Emonet, P. Germain, A. Habrard, E. Morvant, et V. Zantedeschi, « Intérêt Des Bornes Désintégrées Pour La Généralisation Avec Des Mesures de Complexité », in *CAp 2022*, 2022.

[18] K. Bascol, R. Emonet, E. Fromont, et J.-M. Odobez, « Unsupervised Interpretable Pattern Discovery in Time Series Using Autoencoders », in *The Joint IAPR International Workshops on Structural and Syntactic Pattern Recognition (SSPR 2016)*, 2016.

[19] K. Bascol, R. Emonet, E. Fromont, et R. Debusschere, « Improving Chairlift Security with Deep Learning », in *International Symposium on Intelligent Data Analysis*, 2017.

[20] K. Bascol, R. Emonet, E. Fromont, A. Habrard, G. Metzler, et M. Sebban, « CONE : Un Algorithme d'optimisation de La F-Mesure Par Pondération Des Erreurs de Classification », in *CAp 2018*, 2018.

[21] K. Bascol, R. Emonet, E. Fromont, A. Habrard, G. Metzler, et M. Sebban, « From Cost-Sensitive to Tight F-measure Bounds », in *AISTATS 2019*, 2019.

[22] K. Bascol, R. Emonet, et E. Fromont, « Improving Domain Adaptation By Source Selection », in *IEEE International Conference on Image Processing*, 2019.

[23] Y. Wang, R. Emonet, E. Fromont, S. Malinowski, E. Menager, L. Mosser, et R. Tavenard, « Classification de Séries Temporelles Basée Sur Des "shapelets" Interprétables Par Réseaux de Neurones Antagonistes », in *CAp 2019 - Conférence Sur l'Apprentissage Automatique*, 2019. https://hal.archives-ouvertes.fr/hal-02268004

[24] Y. Wang, R. Emonet, E. Fromont, S. Malinowski, E. Menager, L. Mosser, et R. Tavenard, « Learning Interpretable Shapelets for Time Series Classification through Adversarial Regularization », *arXiv preprint arXiv:1906.00917*, 2019, https://arxiv.org/abs/1906.00917

[25] Y. Wang, R. Emonet, E. Fromont, S. Malinowski, et R. Tavenard, « Adversarial Regularization for Explainable-by-Design Time Series Classification », in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020.

[26] T. Kerdoncuff et R. Emonet, « IoU Is Not Submodular », *arXiv preprint arXiv:1809.00593*, 2018, https://arxiv.org/abs/1809.00593

[27] S. Dhouib, I. Redko, T. Kerdoncuff, R. Emonet, et M. Sebban, « A Swiss Army Knife for Minimax Optimal Transport », in *Thirty-Seventh International Conference on Machine Learning*, 2020. https://hal.archives-ouvertes.fr/hal-02900712

[28] T. Kerdoncuff, R. Emonet, et M. Sebban, « Metric Learning in Optimal Transport for Domain Adaptation », in *International Joint Conference on Artificial Intelligence*, 2021. https://hal-ujm.archives-ouvertes.fr/ujm-02611800

[29] T. Kerdoncuff, R. Emonet, et M. Sebban, « Sampled Gromov Wasserstein », *Machine Learning*, 2021, https://hal.science/hal-03232509v2/preview/Sampled_Gromov_Wasserstein_2.pdf

[30] T. Kerdoncuff, M. Perrot, R. Emonet, et M. Sebban, « Optimal Tensor Transport », in *Proceedings (AAAI Artificial Intelligence Conference)*, 2022.

[31] E. Brandao, J.-P. Colombier, S. Duffner, R. Emonet, *et al.*, « Learning PDE to Model Self-Organization of Matter », *Entropy*, 2022, https://www.mdpi.com/1099-4300/24/8/1096

[32] E. Brandao, S. Duffner, R. Emonet, A. Habrard, F. Jacquenet, et M. Sebban, « Is My Neural Net Driven by the MDL Principle? », in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2023. https://hal.science/hal-04231405/

[33] E. Brandao, A. Nakhoul, S. Duffner, R. Emonet, *et al.*, « Learning Complexity to Guide Light-Induced Self-Organized Nanopatterns », *Physical Review Letters*, 2023, https://link.aps.org/doi/10.1103/PhysRevLett.130.226201

[34] J. Varadarajan, R. Emonet, et J.-M. Odobez, « Probabilistic Latent Sequential Motifs: Discovering Temporal Activity Patterns in Video Scenes », in *BMVC*, 2010.

[35] J. Varadarajan, R. Emonet, et J.-M. Odobez, « A Sparsity Constraint for Topic Models - Application to Temporal Activity Mining », in *NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*, 2010.

[36] R. Emonet, J. Varadarajan, et J.-M. Odobez, « Extracting and Locating Temporal Motifs in Video Scenes Using a Hierarchical Non Parametric Bayesian Model », in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. http://www.idiap.ch/~odobez/publications/EmonetVaradarajanOdobez-CVPR-2011.pdf

[37] R. Emonet, J. Varadarajan, et J.-M. Odobez, « Multi-Camera Open Space Human Activity Discovery for Anomaly Detection », in *2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2011.

[38] *Intelligent Video Surveillance Systems (ISTE)*. Wiley-ISTE, 2012. http://www.amazon.com/Intelligent-Video-Surveillance-Systems-ISTE/dp/1848214332%3FSubscriptionId%3D0JYN1NVW651KCA56C102%26tag%3Dtechkie-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D1848214332

[39] *Outils d'analyse Vidéo : Pour Une Pleine Exploitation Des Données de Vidéoprotection*. Hermes Science Publications, 2012. http://www.amazon.co.uk/Outils-danalyse-vid%C3%A9o-exploitation-vid%C3%A9oprotection/dp/274623890X/ref=sr_1_3?s=books

[40] J.-M. Odobez, C. Carincotte, R. Emonet, E. Jouneau, *et al.*, « Unsupervised Activity Analysis and Monitoring Algorithms for Effective Surveillance Systems », 2012.

[41] R. Emonet, J.-M. Odobez, et E. Oberzaucher, « Automatic Discovery Of Recurrent Motion Activities », in *ISHE (Biennial International Conference on Human Ethology)*, 2012. http://media.anthro.univie.ac.at/ishe_conferences/index.php/international/ishe2012/schedConf/program

[42] J. Varadarajan, R. Emonet, et J.-M. Odobez, « Bridging the Past, Present and Future: Modeling Scene Activities From Event Relationships and Global Rules », in *IEEE Conference on Computer Vision and Pattern Recognition, 2012, Providence, Rhode Island, USA*, 2012.

[43] A. Aubert, R. Tavenard, R. Emonet, A. De Lavenne, *et al.*, « Clustering Flood Events from Water Quality Time-Series Using Latent Dirichlet Allocation Model », *Water Resources Research*, 2013,

[44] A. Aubert, R. Tavenard, R. Emonet, S. Malinowski, *et al.*, « Discovering Temporal Patterns in Water Quality Time Series, Focusing on Floods with the LDA Method », in *European Geosciences Union (EGU)*, 2013.

[45] T. Chockalingam, R. Emonet, et J.-M. Odobez, « Localized Anomaly Detection via Hierarchical Integrated Activity Discovery », in *AVSS*, 2013.

[46] R. Tavenard, R. Emonet, et J.-M. Odobez, « Time-Sensitive Topic Models for Action Recognition in Videos », in *IEEE International Conference on Image Processing*, 2013.

[47] J. Varadarajan, R. Emonet, et J.-M. Odobez, « A Sequential Topic Model for Mining Recurrent Activities from Long Term Video Logs », *International Journal of Computer Vision*, 2013.

[48] R. Emonet, J. Varadarajan, et J.-M. Odobez, « Temporal Analysis of Motif Mixtures Using Dirichlet Processes », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[49] R. Emonet, E. Oberzaucher, et J.-M. Odobez, « What to Show? Automatic Stream Selection Among Multiple Sensors », in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2014.

[50] *Practical Applications of Sparse Modeling (Neural Information Processing Series)*. The MIT Press, 2014. http://www.amazon.com/Practical-Applications-Modeling-Information-Processing/dp/0262027720%3FSubscriptionId%3D0JYN1NVW651KCA56C102%26tag%3Dtechkie-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0262027720

[51] J. Varadarajan, R. Emonet, et J.-M. Odobez, « Sparsity in Topic Models », in *Practical Applications of Sparse Modeling (Neural Information Processing Series)*, MIT Press, 2014.

[52] K. W. Mohiuddin, J. Varadarajan, R. Emonet, J. M. Odobez, et P. Moulin, « GPU Accelerated Probabilistic Latent Sequential Motifs for Activity Analysis », in *VISIGRAPP 2018*, 2018.

[53] M. Rußwurm, S. Lefèvre, N. Courty, R. Emonet, M. Körner, et R. Tavenard, « End-to-End Learning for Early Classification of Time Series », *arXiv preprint arXiv:1901.10681*, 2019, https://arxiv.org/abs/1901.10681

[54] M. Rußwurm, N. Courty, R. Emonet, S. Lefèvre, D. Tuia, et R. Tavenard, « End-to-End Learned Early Classification of Time Series for in-Season Crop Type Mapping », *ISPRS Journal of Photogrammetry and Remote Sensing*, 2023.

[55] J. Li, S. Gong, et T. Xiang, « On-the-Fly Global Activity Prediction and Anomaly Detection », in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, 2009.

[56] A. Torralba, A. Willsky, E. Sudderth, et W. Freeman, « Describing Visual Scenes Using Transformed Dirichlet Processes », *NeurIPS*, 2005.

[57] B. Fernando, R. Aljundi, R. Emonet, A. Habrard, M. Sebban, et T. Tuytelaars, « Unsupervised Domain Adaptation Based on Subspace Alignment », in *Domain Adaptation in Computer Vision Applications*, Springer, 2017.

[58] R. Aljundi, R. Emonet, D. Muselet, et M. Sebban, « Landmarks-Based Kernelized Subspace Alignment for Unsupervised Domain Adaptation », in *Computer Vision and Pattern Recognition*, 2015.

[59] B. Fernando, A. Habrard, M. Sebban, et T. Tuytelaars, « Unsupervised Visual Domain Adaptation Using Subspace Alignment », in *Proceedings of the IEEE International Conference on Computer Vision*, 2013. http://www.cv-foundation.org/openaccess/content_iccv_2013/html/Fernando_Unsupervised_Visual_Domain_2013_ICCV_paper.html

[60] G. Csurka, *Domain Adaptation in Computer Vision Applications*. Springer, 2017. http://www.springer.com/gp/book/9783319583464

[61] T. Kerdoncuff, « Contributions to Optimal Transport for Machine Learning: Ground Metric and Generalized Framework », thèse de doctorat, Université de Lyon, 2021. https://hal.science/tel-03534163

[62] N. Bonneel, J. Rabin, G. Peyré, et H. Pfister, « Sliced and Radon Wasserstein Barycenters of Measures », *Journal of Mathematical Imaging and Vision*, 2015, https://doi.org/10.1007/s10851-014-0506-3

[63] R. Viola, R. Emonet, A. Habrard, G. Metzler, S. Riou, et M. Sebban, « A Nearest Neighbor Algorithm for Imbalanced Classification », *International Journal on Artificial Intelligence Tools*, 2021.

[64] R. Viola, R. Emonet, A. Habrard, G. Metzler, et M. Sebban, « MLFP: Un Algorithme d'apprentissage de Métrique Pour La Classification de Données Déséquilibrées », in *Conférence Sur l'Apprentissage Automatique (CAp 2020)*, 2020. https://hal.archives-ouvertes.fr/hal-02868502

[65] R. Viola, R. Emonet, A. Habrard, G. Metzler, et M. Sebban, « Learning from Few Positives: A Provably Accurate Metric Learning Algorithm to Deal with Imbalanced Data », in *IJCAI-PRICAI2020, the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, 2020. https://hal.archives-ouvertes.fr/hal-02611586

[66] R. Viola, R. Emonet, A. Habard, G. Metzler, S. Riou, et M. Sebban, « Une Version Corrigée de l'algorithme Des plus Proches Voisins Pour l'optimisation de La F-mesure Dans Un Contexte Déséquilibré », in *Conférence Sur l'Apprentissage Automatique (CAp 2019)*, 2019. https://hal.archives-ouvertes.fr/hal-02868516

[67] R. Viola, R. Emonet, A. Habrard, G. Metzler, S. Riou, et M. Sebban, « An Adjusted Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data », in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019.

[68] S. Puthiya Parambath, N. Usunier, et Y. Grandvalet, « Optimizing F-measures by Cost-Sensitive Classification », *NeurIPS*, 2014, https://proceedings.neurips.cc/paper/2014/hash/678a1491514b7f1006d605e9161946b1-Abstract.html

[69] A. Sanyal, P. Kumar, P. Kar, S. Chawla, et F. Sebastiani, « Optimizing Non-Decomposable Measures with Deep Networks », *Machine Learning*, 2018, http://link.springer.com/10.1007/s10994-018-5736-y

[70] L. Gautheron, A. Habrard, E. Morvant, et M. Sebban, « Metric Learning from Imbalanced Data with Generalization Guarantees », *Pattern Recognition Letters*, 2020, https://www.sciencedirect.com/science/article/pii/S0167865520300866

[71] J. Cerquides, R. Emonet, G. Picard, et J. A. Rodríguez-Aguilar, « Improving Max-Sum through Decimation to Solve Loopy Distributed Constraint Optimization Problems », in *International Workshop on Optimisation in Multi-Agent Systems (OptMAS@AAMAS 2018)*, 2018.

[72] P. Viallard, R. Emonet, P. Germain, A. Habrard, et E. Morvant, « Interpreting Neural Networks as Majority Votes through the PAC-Bayesian Theory », in *NeurIPS 2019 Workshop on Machine Learning with Guarantees*, 2019.

[73] J. Cerquides, J. A. Rodríguez-Aguilar, R. Emonet, et G. Picard, « Solving Highly Cyclic Distributed Optimization Problems without Busting the Bank: A Decimation-Based Approach », *Logic Journal of the IGPL*, 2021.

[74] K. Kawaguchi, L. P. Kaelbling, et Y. Bengio, « Generalization in Deep Learning », *arXiv preprint arXiv:1710.05468*, 2017, https://arxiv.org/abs/1710.05468

[75] K. (Katayoun). Farrahi, R. Emonet, et A. Ferscha, « Socio-Technical Network Analysis from Wearable Interactions », in *International Symposium on Wearable Computers (ISWC)*, 2012.

[76] K. Farrahi, R. Emonet, et M. Cebrian, « Epidemic Contact Tracing via Communication Traces », *PLoS ONE*, 2014, http://dx.doi.org/10.1371%2Fjournal.pone.0095133

[77] T. Kekec, R. Emonet, E. Fromont, A. Trémeau, et C. Wolf, « Contextually Constrained Deep Networks for Scene Labeling », in *Proceedings of the British Machine Vision Conference, 2014*, 2014.

[78] T. Kekeç, R. Emonet, E. Fromont, A. Trémeau, et C. Wolf, « Prise En Compte Du Contexte Pour Contraindre Les Réseaux Profonds: Application à l'Étiquetage de Scènes », in *Conférence d'Apprentissage, CAP 2014, Saint-Etienne*, 2014.

[79] A. López-Méndez, C. Westling, R. Emonet, M. Easteal, L. Lavia, H. Witchel, et J.-M. Odobez, « Automated Bobbing and Phase Analysis to Measure Walking Entrainment to Music », in *IEEE International Conference on Image Processing*, 2014.

[80] K. Farrahi, R. Emonet, et M. Cebrian, « Predicting a Community's Flu Dynamics with Mobile Phone Data », in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2015.

[81] G. A. Devenyi, R. Emonet, R. M. Harris, K. L. Hertweck, D. Irving, I. Milligan, et G. Wilson, « Ten Simple Rules for Collaborative Lesson Development », *PLoS Computational Biology*, 2018.

[82] R. Ravaille, O. Alata, et R. Emonet, « Extraction d'un Noyau Non Stationnaire de Processus Gaussien à l'aide Des Ondelettes », 2019.

[83] A. C. Rusu, R. Emonet, et K. Farrahi, « Modelling Digital and Manual Contact Tracing for COVID-19. Are Low Uptakes and Missed Contacts Deal-Breakers? », *PloS one*, 2021.

[84] M. Waniek, P. Holme, K. Farrahi, R. Emonet, M. Cebrian, et T. Rahwan, « Trading Contact Tracing Efficiency for Finding Patient Zero », *Scientific reports*, 2022.

[85] J. Li et L. Lin, « Optimal Transport With Relaxed Marginal Constraints », *IEEE Access*, 2021, https://ieeexplore.ieee.org/iel7/6287639/9312710/09400835.pdf

[86] I. Redko, N. Courty, R. Flamary, et D. Tuia, « Optimal Transport for Multi-source Domain Adaptation under Target Shift », in *AISTATS*, 2019. https://proceedings.mlr.press/v89/redko19a.html

[87] T. Vayer, R. Flamary, N. Courty, R. Tavenard, et L. Chapel, « Sliced Gromov-Wasserstein », in *NeurIPS*, 2019. https://proceedings.neurips.cc/paper/2019/hash/a9cc6694dc40736d7a2ec018ea566113-Abstract.html

[88] M. Li, J. Yu, H. Xu, et C. Meng, « Efficient Approximation of Gromov-Wasserstein Distance Using Importance Sparsification ». 10.48550/arXiv.2205.13573. http://arxiv.org/abs/2205.13573

[89] H. Xu, D. Luo, et L. Carin, « Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching », in *NeurIPS*, 2019. https://proceedings.neurips.cc/paper/2019/hash/6e62a992c676f611616097dbea8ea030-Abstract.html

[90] Q. Mérigot, « A Multiscale Approach to Optimal Transport », *Computer Graphics Forum*, 2011, https://onlinelibrary.wiley.com/doi/10.1111/j.1467-8659.2011.02032.x

[91] A. Mallasto, M. Heinonen, et S. Kaski, « Bayesian Inference for Optimal Transport with Stochastic Cost », in *ACML*, 2021. https://proceedings.mlr.press/v157/mallasto21a.html

[92] A. Barbe, M. Sebban, P. Gonçalves, P. Borgnat, et R. Gribonval, « Graph Diffusion Wasserstein Distances », in *Machine Learning and Knowledge Discovery in Databases*, 2021.

[93] G. Mena et J. Niles-Weed, « Statistical Bounds for Entropic Optimal Transport: Sample Complexity and the Central Limit Theorem », in *NeurIPS*, 2019. https://proceedings.neurips.cc/paper/2019/hash/5acdc9ca5d99ae66afdfe1eea0e3b26b-Abstract.html

[94] A. Genevay, L. Chizat, F. Bach, M. Cuturi, et G. Peyré, « Sample Complexity of Sinkhorn Divergences », in *AISTATS*, 2019. https://proceedings.mlr.press/v89/genevay19a.html

[95] R. Ohana, K. Nadjahi, A. Rakotomamonjy, et L. Ralaivola, « Shedding a PAC-Bayesian Light on Adaptive Sliced-Wasserstein Distances », in *International Conference on Machine Learning*, 2023. https://proceedings.mlr.press/v202/ohana23a.html

[96] Z. Lin, Y.-F. Wu, S. V. Peri, W. Sun, *et al.*, « SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition », in *ICLR*, 2022. https://openreview.net/forum?id=rkl03ySYDH

[97] W. Zhu, Y. Shen, M. Liu, et L. P. Aguirre Sanchez, « GMAIR: Unsupervised Object Detection Based on Spatial Attention and Gaussian Mixture Model », *Computational Intelligence and Neuroscience*, 2022, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9313923/

[98] L. C. Evans, « Partial Differential Equations and Monge-Kantorovich Mass Transfer », *Current Developments in Mathematics*, 1997, https://www.intlpress.com/site/pub/pages/journals/items/cdm/content/vols/1997/0001/a002/index.php

[99] Y. Brenier, « Extended Monge-Kantorovich Theory », in *Optimal Transportation and Applications: Lectures given at the C.I.M.E. Summer School, Held in Martina Franca, Italy, September 2-8, 2001*, L. Ambrosio, L. A. Caffarelli, Y. Brenier, G. Buttazzo, C. Villani, et S. Salsa, Éd. Berlin, Heidelberg: Springer, 2003. https://doi.org/10.1007/978-3-540-44857-0_4

[100] P. Mokrov, A. Korotin, L. Li, A. Genevay, J. M. Solomon, et E. Burnaev, « Large-Scale Wasserstein Gradient Flows », in *NeurIPS*, 2021. https://proceedings.neurips.cc/paper/2021/hash/810dfbbebb17302018ae903e9cb7a483-Abstract.html

[101] V. Khrulkov, G. Ryzhakov, A. Chertkov, et I. Oseledets, « Understanding DDPM Latent Codes Through Optimal Transport », in *The Eleventh International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=6PIrhAx1j4i

[102] E. Clerico, G. Deligiannidis, B. Guedj, et A. Doucet, « A PAC-Bayes Bound for Deterministic Classifiers ». 10.48550/arXiv.2209.02525. http://arxiv.org/abs/2209.02525