Recherche de vecteurs caractéristiques robustes via des approches faiblement supervisées : Application à la classification de Supernovae.

Marc Chaumont[1], Johanna Pasquet[4], Jérôme Pasquet[2,3], Fréderic Comby[1], Nancy Rodriguez[1], Dominique Fouchez[4]

1. LIRMM : 161 rue Ada, 34392 Montpellier. marc.chaumont@lirmm.fr, frederic.comby@lirmm.fr, nancy.rodriguez@lirmm.fr
2. TETIS : Maison de la télédection, 500 rue Jean-François Breton 34093 Montpellier. jerome.pasquet@univ-montp3.fr
3. AMIS : Université Paul Valéry, Montpellier, France
4. CPPM : Campus Universitaire de Luminy, 163 Avenue de Luminy, 13009 Marseille Johanna Pasquet : pasquet@cppm.in2p3.fr ï Dominique Fouchez : fouchez@cppm.in2p3.fr

Mots clés : *Traitement du signal, Classification, Clustering, Machine-learning, Deep-Learning, Cosmologie, Supernovae.*

Fig.1: Représentation visuelle de ce à quoi pourrait ressembler une super nova (https://www.supernovae.net/)

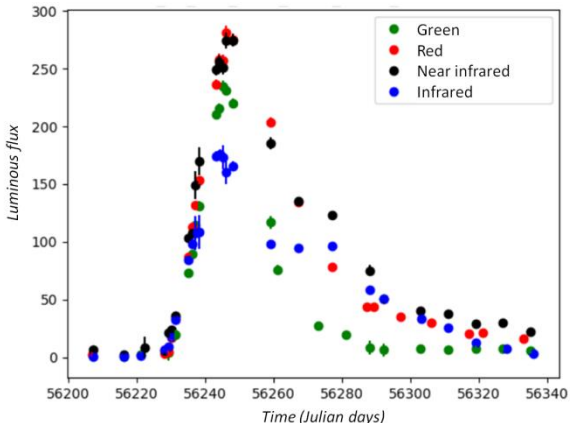

Fig.2 : Courbes de lumière simulée d'une supernovae

La cosmologie tend à comprendre lôorigine, la composition et lôévolution de notre Univers. En 2011, la découverte par Saul Perlmutter, Adam Riess et Brian P. Schmidt de lôexpansion de lôUnivers marque un tournant majeur dans cette science. Cette découverte fut réalisée via lôobservation de Supernovae de type Ia (voir Fig. 1). Les Supernovae sont des étoiles en fin de vie qui explosent, en émettant une très grande quantité de lumière. La détection et lôanalyse de ces astres est donc crucial pour améliorer nos connaissances du modèle actuel de lôunivers.

Les futurs grands relevés astronomiques, tel que le Large Synoptic Survey Telescope (LSST), dont le début des observations commencera en 2022, devrait observer des centaines de milliers de supernovae pendant 10 ans. Afin dôidentifier chaque type dôobjet, et alors de trouver des supernovae Ia, la méthode qui existe actuellement consiste à effectuer un suivi spectral de chaque objet pour en analyser les différentes raies. Cette technique permet dôobtenir un résultat fiable si lôobjet est suffisamment lumineux mais elle est très coûteuse en temps dôobservation et ne peut pas être déployée sur la totalité des astres. Il convient alors dôutiliser seulement lôinformation photométrique, côest à dire la réponse en amplitude dôun objet céleste dans quelques bandes spectrales (voir Fig. 2). Cependant pour construire une base dôapprentissage pour les algorithmes de classification, il convient de disposer dôun petit échantillon dont le spectre a été mesuré. Cet échantillon est souvent petit et biaisé car il contient principalement les objets les plus brillants. Il y a donc un décalage entre les distributions de flux lumineux des bases d'entraînement et de test. Ce décalage appelé "mismatch" entre la base

d'apprentissage et la base cible [6, 7] est un problème connu en apprentissage automatique que nous proposons dôétudier dans ce stage. Les observations sont réalisées sur une période de temps donnée. On peut alors construire une série temporelle, appelée « courbe de lumière » (voir Fig. 2) qui représente les valeurs de lôamplitude dôun objet dans une bande spécifique au cours du temps.

Afin dôanalyser ces données plusieurs approches ont mises en avant des algorithmes dôapprentissage supervisé [3,4,5] qui apprenne un modèle sur lôéchantillon d'entraînement petit et biaisé puis est testé sur les données issues de la photométrie. Dans ce travail nous souhaitons nous affranchir du problème de classification pour nous focaliser sur la représentation des données. Lôobjectif est alors de proposer une architecture non supervisée [1] ou semi-supervisé [2] permettant lôextraction de caractéristiques robustes et discriminantes. Dans un second temps nous souhaitons mettre en place des algorithmes de clustering afin de regrouper ces caractéristiques par affinité.

Ce travail sera contextualisé autour des données simulées de LSST fournies par le challenge Kaggle PLAsTiCC-2018. Afin de vérifier la cohérence des clusters obtenus nous superposerons les différents clusters obtenus avec la vérité des classes connues.

Notons que ce travail est la suite d'une première collaboration entre Montpellier (LIRMM/TETIS) et Marseille (CPPM : Centre de Physique des Particules de Marseille), dont l'article suivant est issu [8].

Profil recherché : Master (M2) ou Ecole dôIngénieur (3ème année) ayant une bonne maitrise de la programmation (C++, Python, Tensorflow), et une ou plusieurs connaissances en traitement des images, fouille de données / indexation / classification, architecture des machines/installation d'OS, Anglais écrit scientifique. **Aucune connaissance en astrophysique nôest requise.**

Modalité de candidature : Envoyez un CV, une lettre de motivation ainsi que votre relevé de notes de M1 le plus tôt possible. Après pré-sélection des candidatures, des entretiens téléphoniques ou en personne seront planifiés.

Contacts : Marc Chaumont (*marc.chaumont@lirmm.fr*) / Jérôme Pasquet (*jerome.pasquet@univ-montp3.fr*)

Lieu du stage : Le stage se déroulera au LIRMM équipe ICAR (campus St Priest), Montpellier, France.

Période du stage : 1er semestre 2018 (5-6 mois).

Gratification de stage : environ 550ú mois.

Bibliographie:

[1] “Autoencoders, Unsupervised Learning, and Deep Architectures”, Pierre Baldi, Proceedings of ICML Workshop on Unsupervised and Transfer Learning 2012

[2] Dimensionality Reduction by Learning an Invariant Mapping - Raia Hadsell, Sumit Chopra, Yann LeCun - CVPR 2006

[3] “Deep learning approach for classifying, detecting and predicting photometric redshifts of quasars in the Sloan Digital Sky Survey stripe 82”, J. Pasquet-Itam and J. Pasquet, A&A 2018

[4] “PELICAN: deeP architecturE for the LIght Curve ANalysis ”, J. Pasquet-Itam, J. Pasquet, M. Chaumont, D. Fouchez, A&A 20189 (en cours d'évaluation)

[5] “”Deep Recurrent Neural Networks for Supernovae Classification”,Tom Charnock, Adam Moss, A&A 2016

[6] "Deep Learning for Imbalanced Multimedia Data Classification" Yilin Yan, Min Chen, MeiLing Shyu, and Shu-Ching Chen, Multimedia (ISM), 2015 IEEE International Symposium on

[7] " The Impact of Imbalanced Training Data for Convolutional Neural Networks", PAULINA HENSMAN AND DAVID MASKO, Degree Project in Computer Science

[8] " A CNN adapted to time series for the classification of Supernovae ", Anthony Brunel, Johanna Pasquet, Jérôme Pasquet, Nancy Rodriguez, Frédéric Comby, Dominique Fouchez, Marc Chaumont, EI'2019, in Proceedings of Color Imaging XXIV: Displaying, Processing, Hardcopy, and Applications - joint sessions dealing with color in astronomy / astrophysics, Part of IS&T International Symposium on Electronic Imaging, Burlingame (suburb of San Francisco), California USA, 13 - 17 January, 2019. The CNN is downloadable there: https://github.com/Anzzy30/SupernovaeClassification