

Classification automatique des différents type de Supernovae à l'aide du Machine Learning

Sujet de M2 Recherche

Marc Chaumont¹, Johanna Pasquet², Nancy Rodriguez¹, Dominique Fouchez², Jérôme Pasquet³, Fabrice Feinstein²

¹ LIRMM :161 rue Ada, 34392 Montpellier cedex 5 - France

Tel : +33 4.67.14.97.59, Marc.Chaumont@lirmm.fr, Nancy.Rodriguez@lirmm.fr

² CPPM : Campus Universitaire de Luminy, 163 Avenue de Luminy, 13009 Marseille – France

Tel : +33 4 91 82 72 62, pasquet@cppm.in2p3.fr

³ LSIS :Polytech Marseille Sud LSIS UMR 7296 Campus de Luminy Case 925, 13288 Marseille -France
jerome.pasquet@lisis.org

Mots clés : *Traitement du signal, classification, machine learning, deep Learning, Cosmologie, Supernovae*

La cosmologie tend à comprendre l'origine, la composition et l'évolution de notre Univers. La découverte de l'accélération de l'expansion de l'Univers marque un tournant majeur. Cette découverte qui a permis à Saul Perlmutter, Adam Riess et Brian P. Schmidt d'obtenir le prix Nobel en 2011, a été réalisée en observant des Supernovae de type Ia. Les Supernovae sont des étoiles en fin de vie qui explosent, en émettant une très grande quantité de lumière. Selon le mécanisme d'explosion on distingue deux grand types de Supernovae. Les Supernovae de type **Ia** sont caractérisées par une explosion thermonucléaire, au cours de laquelle une naine blanche explose à la suite d'une accrétion de matière d'une étoile compagnon. Le mécanisme d'explosion des étoiles plus massives (Supernovae **II**, **Ib**, **Ic**) mettent en jeu un effondrement gravitationnel, au cours de laquelle le coeur de fer de l'étoile s'effondre sur lui même.



Les Supernovae de type **Ia** sont particulièrement intéressantes car le mécanisme d'explosion est semblable pour l'ensemble de ces étoiles. Ces explosions étant particulièrement lumineuses, il est possible de les voir à de très grande distance, ce qui permet de sonder l'Univers lointain et de comprendre la nature de l'énergie noire.

La méthode qui permet d'identifier une Supernovae consiste à analyser les raies spectrales de cet objet. Avec l'arrivée des grands relevés astronomiques, tel que le Large Synoptic Survey Telescope en 2022, qui mesurera la luminosité d'environ 37 milliards d'étoiles et de galaxies pendant 10 ans, il est impossible de réaliser le spectre de chaque objet (faire le spectre d'un objet astrophysique requiert que cet objet soit suffisamment lumineux donc proche de nous et un temps de pose important). Il convient alors d'utiliser seulement l'information photométrique, c'est à dire la réponse en amplitude d'un objet céleste dans quelques bandes spectrales (par exemple le bleu, le visible, le proche infra-rouge). Les observations sont réalisées sur une période de temps donnée. On peut alors construire une série temporelle, appelée « courbe de lumière » qui représente les valeurs de l'amplitude d'un objet dans une bande spécifique au cours du temps.

Pour automatiser le processus de classification de chacun des objets observés, et en particulier l'identification des différents types de Supernovae à partir des "courbes de lumières", il faut mettre en place des algorithmes de classification efficaces. L'approche que nous envisageons consiste à utiliser une méthodologie d'apprentissage par deep learning [5] [6].

Il y a cependant deux limitations majeures à utiliser des méthodes d'apprentissage pour cette étude :
- La première difficulté est que la base d'apprentissage annotée (les exemples de courbes de lumière) ne contient que des objets célestes lumineux et donc "proches" de nous. En effet, la vérité sur les objets présents dans la base d'apprentissage est obtenue à partir de mesures spectrales. La

base d'apprentissage ne contient donc pas d'objets peu lumineux et éloignés, alors que la base à tester contiendra à la fois des objets proches et éloignés. Ce problème de représentativité entre la base d'entraînement et la base de test crée un problème de *mismatch* qui diminue fortement les performances des algorithmes.

- La deuxième difficulté est que la proportion des différents types d'objets célestes dans la base d'apprentissage n'est pas la même que celle dans la base de test [7][8].

Pour se préparer aux futures données, un challenge avait été lancé en 2010 ([1], [2]) dans lequel des courbes de lumière ont été synthétisées à l'aide de simulations basées sur des données réelles. L'objectif était de classer environ 20 000 courbes de lumière de différents types de Supernovae. Pour cela, un échantillon d'entraînement était fourni et comportait environ 1000 courbes de lumière. Cet échantillon est biaisé comme mentionné précédemment et est également asymétrique, puisque les Supernovae de type Ia y sont sur-représentées par rapport aux autres types de Supernovae. Plusieurs travaux ont été réalisés mais beaucoup d'entre eux ont contournés le problème de *mismatch* en reconstruisant un lot d'entraînement plus représentatif ([3], [4]). Un autre challenge devrait être lancé en fin d'année 2017 pour une durée de 6 mois avec pour objectif la classification des Supernovae mais aussi d'autres objets variables tels que des quasars, des étoiles variables...

Le but de ce stage est de développer une méthode de classification basée sur du Deep Learning pour classer les Supernovae de type Ia dans le cadre du challenge photométrique qui aura débuté depuis peu. Des travaux préliminaires, en cours de publication ont montré que l'utilisation du Deep Learning donnait des résultats proches voir meilleurs que l'état de l'art. Les problèmes de mismatch, de quantité d'exemples, d'équilibrage des bases de données, ainsi que d'utilisation du Deep Learning seront à explorer durant ce stage. Il sera également possible d'explorer d'autres algorithmes de classification impliquant une étape d'extraction de caractéristiques. L'étudiant disposera du travail déjà réalisé dans le cadre du challenge de 2010.

Pré requis souhaitables : programmation C/C++ et/ou Python, anglais écrit scientifique.
Aucune connaissance en astrophysique n'est requise.

Conditions de stage :

Durée : 5 à 6 mois

Indemnités : environs à 550 € / mois.

Le stage se déroulera au LIRMM (campus St Priest), Montpellier, France.

Références

[1] Kessler R., Conley A., Jha S., Kuhlmann S., 2010a, arXiv preprint arXiv:1001.5210

[2] Kessler, R., Bassett, B., Belov, P., et al. 2010, PASP, 122, 1415

[3] Ishida E. E. O., de Souza R. S., 2013, MNRAS, 430, 509

[4] Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, ArXiv e-prints, arXiv:1603.00882

[5] "Densely Connected Convolutional Networks (CVPR 2017 Best Paper Award)", Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Weinberger, IEEE Conference on Computer Vision and Pattern Recognition, CVPR'2017

[6] "CNN Architectures for Large-Scale Audio Classification" Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, Kevin Wilson, International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2017).

- [7] "Deep Learning for Imbalanced Multimedia Data Classification" Yilin Yan, Min Chen, Mei-Ling Shyu, and Shu-Ching Chen, Multimedia (ISM), 2015 IEEE International Symposium on
- [8] " The Impact of Imbalanced Training Data for Convolutional Neural Networks", PAULINA HENSMAN AND DAVID MASKO, Degree Project in Computer Science,
https://www.kth.se/social/files/588617ebf2765401cfcc478c/PHensmanDMasko_dkand15.pdf