Sur la classification d'étoiles en fonction de leur spectre d'absorption par apprentissage automatique

Patrice Béchard
Département d'informatique
et de recherche opérationelle
Université de Montréal
Montréal, QC H3T 1J4
patrice.bechard@umontreal.ca

Jean-Pascal Guévin
Département de mathématiques
et de statistique
Université de Montréal
Montréal, QC H3T 1J4
jean-pascal.guevin@umontreal.ca

IFT6390 - Fondements de l'apprentissage machine - 18 décembre 2017

Abstract

1 Introduction

L'aspect principal du projet consiste en la classification d'étoiles en fonction de leur type spectral. L'exploration de notre univers observable a en effet amené les astrophysiciens à observer et à catégoriser des centaines de milliers d'étoiles en fonction notamment de leur taille, de leur masse, de leur température et de leur composition. Ce processus de classification se fait, entre autre, à partir du spectre d'absorption des étoiles, c'est-à-dire une mesure de l'intensité du spectre électromagnétique émis par celles-ci en fonction de la longueur d'onde.

2 Méthodes

Les données utilisées pour les spectres d'étoiles proviennent de la base de données *Sloan Digital Sky Survey* (SDSS) *Science Archive Server* (SAS) donnant gratuitement accès aux observations faites par différents télescopes. Il est évidemment nécessaire de traiter les spectres obtenus par le biais du SDSS, ceux-ci étant généralement très bruités. Un processus lissage et de normalisation permet d'en extraire l'information pertinente en éliminant le plus possible le bruit et en ne conservant que ce qui semble correspondre à des tendances plus globales. De plus, chaque spectre a été tronqué de sorte que seul le flux correspondant aux log-longueurs d'onde entre 3.65 et 3.80 (correspondant aux longueurs d'onde entre ≈ 398.1 nm jusqu'à ≈ 707.9 nm, ce qui représente le spectre de lumière visible). Finalement, une interpolation linéaire de points a permis de diminuer le nombre de traits caractéristiques à 1000, ce que les algorithmes peuvent manipuler sans problème. Un échantillon de 10000 étoiles pour chaque 6 types spectral différent utilisé (A, F, G, K, M, WD) a été traité. Puisque ces spectres sont les seules entrées des algorithmes essayés, le traitement des données a un impact majeur sur les résultats. La figure 1 présente un exemple d'un spectre d'étoile avant et après avoir traiter les données.

Une première validation des algorithmes est effectuée par le biais d'une tâche connexe, soit la classification d'électrocardiogrammes selon l'état de santé du patient duquel il provient. Les électrocardiogrammes (ECG) étant fort semblables dans leurs formes à des spectres d'étoiles (tous deux étant des données corrélées dans l'espace en 1 dimension), ceci permet un premier ajustement des algorithmes en plus de nous initier au fonctionnement de ceux-ci dans le contexte d'une analyse spectroscopique. Les électrocardiogrammes qui utilisés proviennent du *PhysioNet/Computing in Cardiology Challenge 2017* et ont l'avantage d'être plus simple à analyser, puisqu'ils sont plus lisses et moins bruités. Une séquence de 10 secondes a été conservée pour chaque ECG et les séquences

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

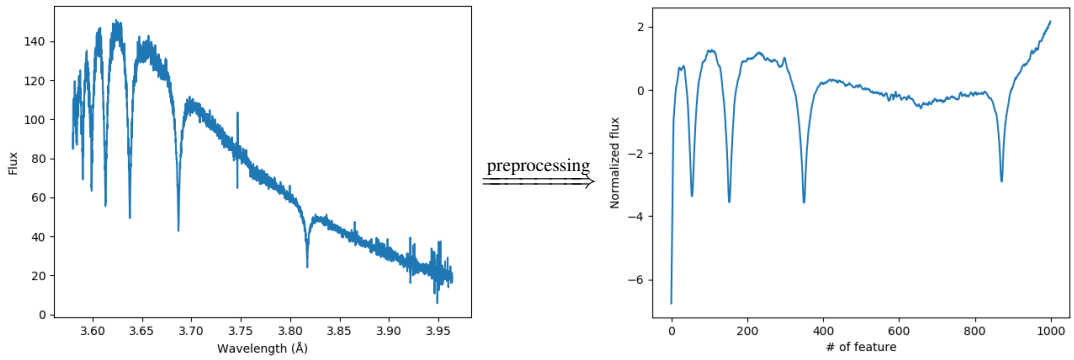

FIGURE 1 – Effect of preprocessing on raw star spectras

ont été classées en 2 catégories, soit un patient en santé (5050 exemples), soit un patient avec une arythmie ou un autre problème cardiaque (3478 exemples). Le nombre de traits caractéristiques pour cet ensemble de données est de 500 pour chaque exemple.

Trois familles d'algorithmes seront étudiées dans le cadre de ce projet. Tout d'abord, nous utiliserons des réseaux de neurones multicouches de type perceptron. L'usage de librairies telles Keras ou TensorFlow rend très simple l'implémentation de ce type d'algorithme. Le réseau créé prendra en entrée un spectre traité (centré, réduit et lissé) et retournera en sortie la classification du spectre selon un encodage *onehot*. Le nombre ainsi que la taille des couches cachées sont des hyperparamètres qu'il faudra déterminer.

Ensuite, puisque les points formant un spectre sont corrélés entre eux, les réseaux de neurones convolutifs sont une approche qui semble prometteuse. Le réseau créé calculera des convolutions unidimensionnelles à partir des données. Le nombre de convolutions, de filtres ainsi que le type de *pooling* (max, moyen, etc.) souhaités restent à déterminer.

Enfin, un arbre de décision sera construit afin de classer les spectres. Cette approche est similaire à celle conduite par Ball et al. (2006), qui visait à classifier des objets célestes en catégories plus grandes (étoile, galaxie ou QSO) à l'aide d'arbres de décision .

3 Résultats

4 Discussion

5 Répartition et remerciements

Références

Ball, N. M., Brunner, R. J., Myers, A. D., and Tcheng, D. (2006). Robust machine learning applied to astronomical data sets. i. star-galaxy classification of the sloan digital sky survey dr3 using decision trees. *The Astrophysical Journal*, 650(1) :497.