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 - 19 décembre 2017

Abstract

1 Introduction

L'aspect principal du projet consiste en la classification d'étoiles en fonction de leur type spectral à l'aide d'algorithmes d'apprentissage. 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. Pour la validation des algorithmes, des données d'électrocardiogramme, étant aussi des données corrélées en 1 dimension, seront utilisées. Les algorithmes d'apprentissage utilisés pour effectuer la classification sont les réseaux de neurones de type MLP, les réseaux de neurones convolutifs ainsi que les machines à vecteur de support. Les bases de données ainsi que les algorithmes d'apprentissages utilisés sont présentés en détails à la section 2 et les résultats obtenus sont présentés à la section 3. Les codes et les figures présentées pour l'ensemble du projet sont disponibles en ligne sur GitHub : `https://github.com/patricebechard/Machine_learning_IFT6390`.

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 électro-

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

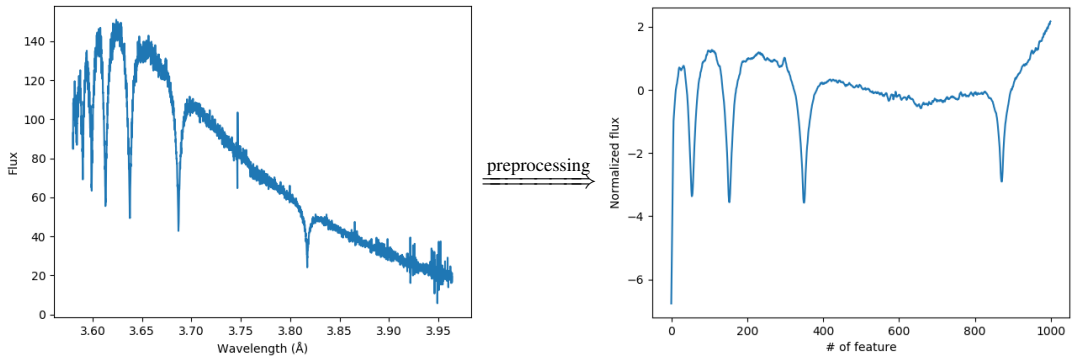

FIGURE 1 – Effet du prétraitement des spectres d'étoiles. À gauche, un exemple de données brutes fournies par le *Sloan Digital Sky Survey* est présenté. À droite, le même exemple a été normalisé et lissé.

cardiogrammes (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 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 300 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*. La méthodologie utilisée pour ajuster les hyperparamètres (nombre de couches cachées, nombre de neurones dans chaque couche, régularisation, taille des lots) consiste à faire un *grid search* sur plusieurs architectures de réseaux ainsi que plusieurs types de régularisation et plusieurs tailles de lots pour trouver une ou plusieurs combinaisons prometteuses en mesurant l'erreur de classification sur l'ensemble de validation. Un processus de *fine tuning* des paramètres suivait ensuite pour améliorer le plus possible l'efficacité de l'algorithme. Finalement, un réentraînement sur l'ensemble d'entraînement ainsi que l'ensemble de validation avait lieu, accompagné d'un test sur l'ensemble de test nous a permis d'obtenir les résultats finaux. Une présentation en détails des résultats obtenus est faite à la section 3. L'utilisation d'un processeur graphique (GPU) nous a permis d'effectuer une panoplie de différentes expériences sans que celles-ci soient trop coûteuses en temps. Des travaux similaires ont été conduits par Bailer-Jones et al. (1998) ainsi que Carricajo et al. (2004) pour les spectres stellaires, et par Shensheng Xu et al. (2017) pour les électrocardiogrammes.

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 sont les hyperparamètres à déterminer. Ce type d'algorithme permet de réduire le nombre de dimensions du problème en plus de prendre en compte des caractéristiques des entrées comme la connectivité locale des traits caractéristiques, tout en introduisant une invariance des traductions locales grâce au *pooling*. Hála (2014) a conduit des travaux similaires pour les spectres d'étoiles. Plusieurs travaux de recherche, notamment par Rajpurkar et al. (2017) ainsi que Zihlmann et al. (2017) se sont penchés sur la classification d'électrocardiogrammes à l'aide de réseaux de neurones convolutifs.

Enfin, les machines à vecteur de support (SVM) seront le dernier type d'algorithme d'apprentissage utilisé pour la classification des spectres stellaires ainsi que des électrocardiogrammes.

TODO : Expliquer SVM

Des travaux similaires ont été conduits par Bu et al. (2014) pour la classification de spectres d'étoile et par Shensheng Xu et al. (2017) pour les électrocardiogrammes.

3 Résultats

3.1 Électrocardiogrammes

TABLE 1 – Résultats pour la classification d'électrocardiogrammes

Algorithme	Erreur de classification (Entraînement)	Erreur de classification (Test)
Arbre de décision	Input terminal	$\sim$100
MLP	Output terminal	$\sim$10
CNN	Cell body	up to 10^6

3.2 Spectres d'étoiles

TABLE 2 – Résultats pour la classification de spectres d'étoiles

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

4 Discussion

5 Répartition et remerciements

Références

Bailer-Jones, C. A., Irwin, M., and Von Hippel, T. (1998). Automated classification of stellar spectra ?ii. two-dimensional classification with neural networks and principal components analysis. *Monthly Notices of the Royal Astronomical Society*, 298(2) :361–377.

Bu, Y., Chen, F., and Pan, J. (2014). Stellar spectral subclasses classification based on isomap and svm. *New Astronomy*, 28 :35–43.

Carricajo, I., Manteiga, M., Rodríguez, A., and Dafonte, C. (2004). Automatic classification of stellar spectra. *Lecture Notes and Essays in Astrophysics*, 1 :153–164.

Hála, P. (2014). Spectral classification using convolutional neural networks. *arXiv preprint arXiv :1412.8341*.

Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., and Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv :1707.01836*.

Shensheng Xu, S., Mak, M.-W., and Cheung, C.-C. (2017). Deep neural networks versus support vector machines for ecg arrhythmia classification. In *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*, pages 127–132. IEEE.

Zihlmann, M., Perekrestenko, D., and Tschannen, M. (2017). Convolutional recurrent neural networks for electrocardiogram classification. *arXiv preprint arXiv :1710.06122*.