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 - 21 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 (CNN) ainsi que les machines à vecteur de support (SVM). 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). Nous avons aplati les spectres en ajustant une courbe de degré 3 aux données et divisé par celle-ci. 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

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

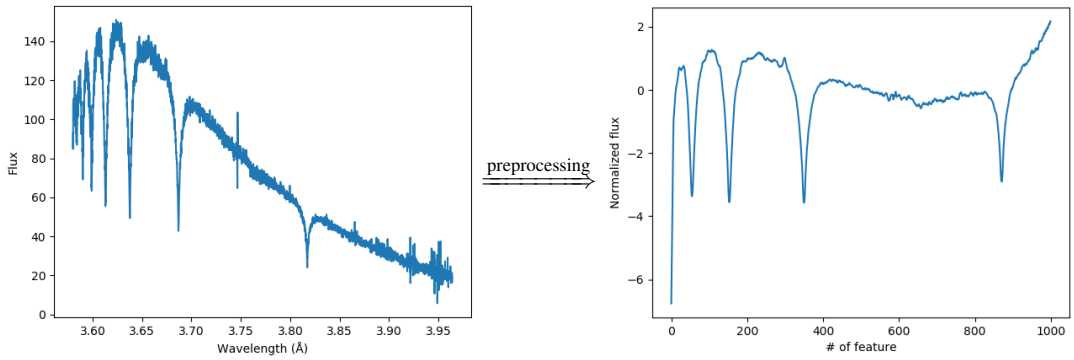

FIGURE 1 – Effet du prétraitement des spectres d'étoiles. À gauche, un exemple de données brutes fournies par le *Sloan Digital Sky Survey* est présenté. À droite, le même exemple a été normalisé et lissé.

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 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 de type perceptron multicouche (MLP). 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 (CNN) 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. Ce genre

d'algorithme a tendance à bien se débrouiller avec des entrées possédant un grand nombre de traits caractéristiques. Le noyau à utiliser sera déterminé lors de la sélection du modèle. De plus, pour la tâche de classification multiclasse des spectres stellaires, le choix du type de comparaison pour les frontières de décision (*one-vs-one* ou *one-vs-rest*) sera à déterminer. L'implémentation des SVM a été conduite à l'aide de la librairie Scikit Learn. 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

Contrairement ce à quoi nous nous attendions, nous avons eu beaucoup plus de problèmes avec les données d'électrocardiogramme qu'avec les données de spectres stellaires. Tout d'abord, pour les SVM, des tests ont été effectués pour déterminer le noyau du SVM à utiliser parmi ceux implémentés par défaut dans Scikit Learn, soit un noyau linéaire(`linear`, $\langle x, x' \rangle$), un noyau à fonction à base radiale (`rbf`, $\exp(-\gamma||x - x'||_2^2)$) avec un facteur γ à déterminer, un noyau polynomial (`poly`, $(\gamma\langle x, x' \rangle + r)^d$) avec des facteurs γ et r à déterminer, ou un noyau sigmoïdal (`sigmoid`, $\tanh(\gamma\langle x, x' \rangle + r)$). Les résultats obtenus pour ces noyaux avec leurs valeurs par défaut sont présentés au tableau 1.

TABLE 1 – Résultats pour la classification d'électrocardiogrammes en fonction du noyau utilisé.

NOYAU	`linear`	`rbf`	`poly`	`sigmoid`
PRÉCISION	55.06%	59.28%	55.06%	52.60%

Étant donné les résultats obtenus, on décide de ne garder que le noyau `rbf` pour simplifier la suite. Ces résultats étant très bas, il est pertinent de se questionner sur la validité du prétraitement des données pour notre problème. Tout d'abord, nous essaierons de réduire la dimensionnalité des données à l'aide d'analyse de composantes principales (PCA) sur les données normalisées, similairement à Polat and Güneş (2007). De plus, puisque les électrocardiogrammes sont des données périodiques, nous essaierons dans un autre cas d'effectuer une transformée de Fourier sur les données normalisées et regarder si l'une de ces méthodes diminue l'erreur de classification. Ce type de prétraitement a été utilisé conjointement avec des réseaux de neurones précédemment, notamment par Gothwal et al. (2011). Les figures **??** et **??** (en annexe) présentent les courbes d'apprentissage obtenues pour un SVM avec noyau RBF.

Finalement, en utilisant un CNN pour faire la détection d'arythmies cardiaques, nous avons essayé plusieurs configurations pour maximiser la classification. En mettant un CNN et un réseau MLP complètement connecté bout-à-bout, nous avons vérifié l'erreur de classification en faisant varier le nombre de couches de convolution et de pooling, ainsi qu'en essayant diverses architectures pour le MLP. Nous nous sommes limités à des tailles de filtres de convolution de largeur 4 et des filtres de pooling de largeur 4. Nous n'avons pas jugé nécessaire de tester le CNN avec les données prétraitées à l'aide de PCA ou d'une transformation de Fourier.

Les résultats finaux de la précision de chaque algorithms sur les électrocardiogrammes est présenté au tableau 2.

TABLE 2 – Résultats pour la classification d'électrocardiogrammes

ALGORITHME	SVM	Réseau de neurones MLP	CNN
PRÉCISION	%	%	%

3.2 Spectres d'étoiles

Les résultats obtenus pour la classification de spectres stellaires ont été beaucoup plus concluants que ceux obtenus pour la classification d'électrocardiogrammes.

TABLE 3 – 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.

Gothwal, H., Kedawat, S., and Kumar, R. (2011). Cardiac arrhythmias detection in an ecg beat signal using fast fourier transform and artificial neural network. *Journal of Biomedical Science and Engineering*, 4(04) :289.

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

Polat, K. and Güneş, S. (2007). Detection of ecg arrhythmia using a differential expert system approach based on principal component analysis and least square support vector machine. *Applied Mathematics and Computation*, 186(1) :898–906.

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*.

Annexe