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 - 15 novembre 2017

1 Proposition du projet

L'aspect principal du projet proposé 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.

Les données utilisées proviendront de la base de données *Sloan Digital Sky Survey* (SDSS) *Science Archive Server* (SAS) donnant gratuitement accès aux observations faites par différents télescopes. Il sera évidemment nécessaire de traiter les spectres obtenus par le biais du SDSS, ceux-ci étant généralement très bruités. Un processus lissage et de normalisation devrait permettre 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. Puisque ces spectres seront les seules entrées des algorithmes qui seront essayés, le traitement des données aura un impact majeur sur les résultats.

Une première validation des algorithmes sera 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 étant fort semblable dans leur forme à des spectres d'étoiles, ceci permettra 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 seront 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. Les fichiers les contenant sont également plus petit, de sorte que leur traitement devrait être plus rapide que pour les spectres d'étoiles. Enfin, ceux-ci sont répartis en seulement deux classes (les patients sains et les patients malades), ce qui devrait faciliter la tâche de classification.

Trois familles d'algorithmes seront étudiées dans le cadre de ce projet :

1. Les 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 encore à déterminer.
2. 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.
3. Enfin, un arbre de décision sera construit afin de classer les spectres.

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