Automating biomedical data science through tree-based pipeline optimization Randal S. Olson, Ryan J. Urbanowicz, Peter C. Andrews, Nicole, A. Lavender, La Creis Kidd et Jason H. Moore

Olfa Lmt

January 2021

1 Introduction

Nous aborderons ici le concept d'optimisation des pipelines à base d'arbres pour automatiser l'une des parties les plus compliqué dans le machine learning, à savoir la conception des pipelines. Nous étudierons ce concept à travers l'outil TPOT, pour Tree-Based Pipeline Optimization Tool.

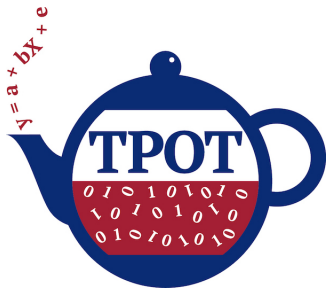

Figure 1: Logo de l'outil TPOT

2 TPOT

TPOT est conçu comme un outil d'aide qui vous donne des idées sur la manière de résoudre un problème d'apprentissage particulier en explorant des configurations de pipeline que vous n'auriez peut-être jamais envisagées, puis qui laisse le réglage fin à des techniques de réglage de paramètres plus contraignants comme

la recherche par grille. TPOT peut construire des pipelines d'apprentissage machine qui atteignent une précision de classification compétitive.

TPOT est open source et bien documenté . Son développement a été mené par des chercheurs de l'Université de Pennsylvanie et est la bibliothèques autoML la plus populaires.

3 Comment fonctionne TPOT ?

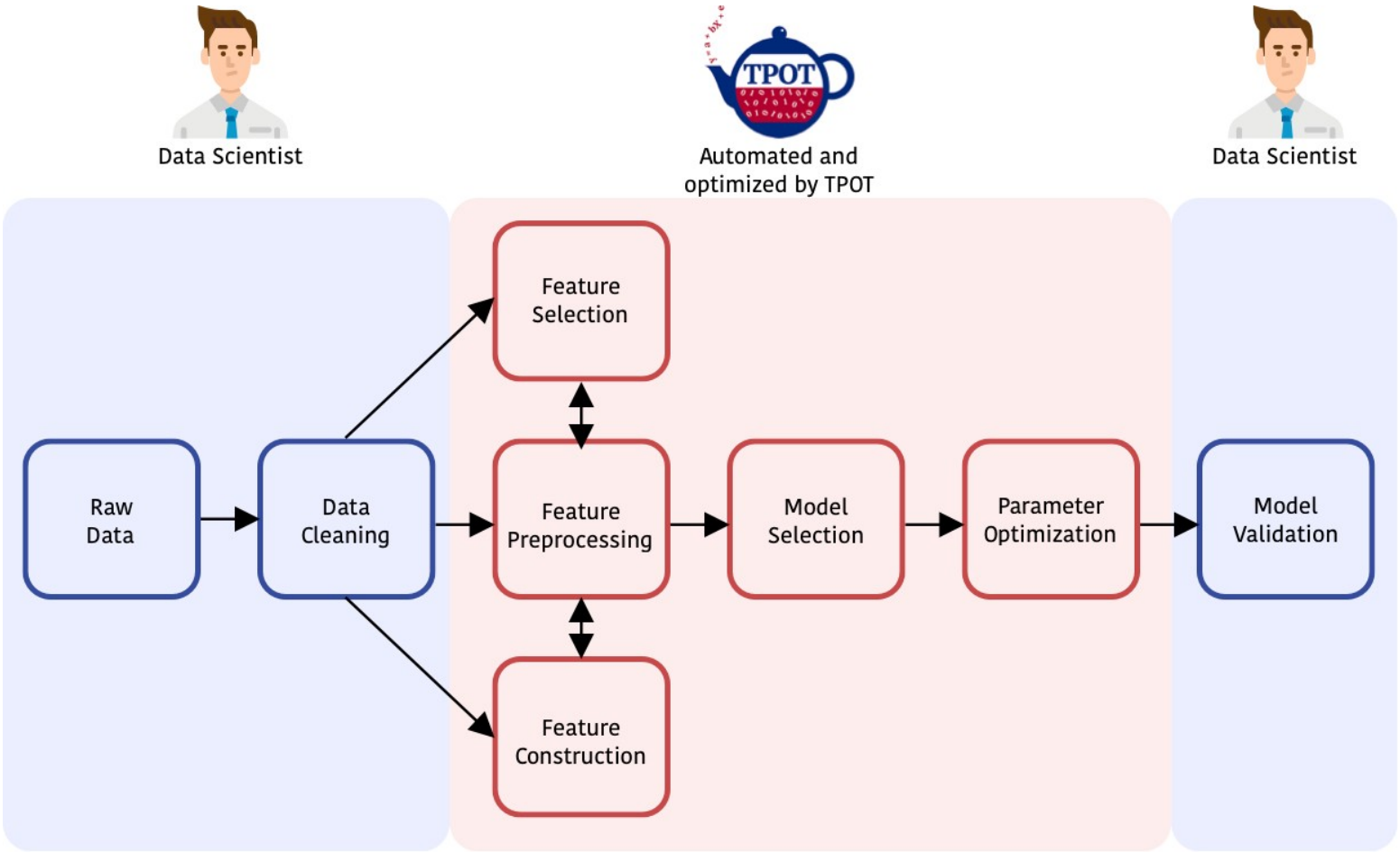

Figure 2: En bleu (à droite et à gauche) sont mis en évidence les composants d'un pipeline traditionnel qui sont adressés par un Data Scientist. En rouge (au milieu) est mis en évidence ce que TPOT couvre

TPOT dispose d'un algorithme de recherche génétique pour trouver les

meilleurs paramètres et ensembles de modèles. On peut aussi le considérer comme un algorithme de sélection naturelle ou d'évolution. TPOT essaie un pipeline, évalue ses performances et modifie aléatoirement certaines parties du pipeline à la recherche d'algorithmes plus performants.

Les algorithmes AutoML ne sont donc pas aussi simples que l'ajustement d'un modèle sur l'ensemble de données. Ils envisagent plutôt plusieurs algorithmes d'apprentissage (comme le modèle linéaires, random forest ou enocre MVS) dans un pipeline avec plusieurs étapes de prétraitement (comme l'imputation des valeurs manquantes, mise à l'échelle, ACP, sélection de caractéristiques, etc. (source : docs TPOT)

Cette puissance du TPOT provient de l'évaluation automatique et efficace de toutes sortes de pipelines possibles car le faire manuellement est lourd et plus lent.

4 Optimisation des pipelines avec des Genetic Algorithms

La programmation génétique (GP) est un type d'algorithme évolutif (EA), un sous-ensemble de l'apprentissage machine. Les EA sont utilisés pour découvrir directement des solutions à des problèmes que les humains ne savent pas résoudre. La nature adaptative des EA peut générer des solutions comparables et souvent meilleures que ce que l'Homme puisse faire.

Chaque opérateur du pipeline est composé d'un ensemble de fonctions où chacune de ces fonctions reçoit un ensemble de paramètres, c'est donc ici que la GP joue un rôle très important. TPOT implémente une technique de calcul évolutive. Plus précisément, TPOT met en œuvre une GP pour faire évoluer séquentiellement les opérateurs de pipeline ainsi que les paramètres des opérateurs afin de maximiser la précision de la classification des pipelines.

Inspirés par l'évolution biologique et ses mécanismes fondamentaux, les GP mettent en œuvre un algorithme qui utilise des mutations aléatoires, des croisements, une fonction fitness (fonction permettant de guider les simulations vers des solutions de conception optimales) et plusieurs générations d'évolution pour résoudre une tâche définie par l'utilisateur. La GP peut être utilisée pour découvrir une relation fonctionnelle entre les caractéristiques des données (régression symbolique), pour regrouper les données en catégories (classification) et pour aider à la conception de circuits électriques, d'antennes et d'algorithmes quantiques. La GP est appliquée au génie logiciel par la synthèse de code, l'amélioration génétique, la correction automatique de bugs, et dans le développement de stratégies de jeu, ... et plus encore.

Les algorithmes génétiques sont inspirés du processus darwinien de sélection naturelle et sont utilisés pour générer des solutions aux problèmes d'optimisation et de recherche en informatique. De manière générale, les algorithmes génétiques ont trois propriétés :

- La sélection : Vous disposez d'une population de solutions possibles à un problème donné et d'une fonction d'aptitude. À chaque itération, vous évaluez comment adapter chaque solution à votre fonction de fitness.

- Croisement : Vous sélectionnez ensuite les solutions les plus adaptées et effectuez un croisement pour créer une nouvelle population.

- Mutation : Vous prenez ces enfants et les faites muter avec des modifications aléatoires et vous répétez le processus jusqu'à ce que vous obteniez la solution la plus adaptée ou la meilleure.

5 Notre avis

Le document ne semble pas très compliqué à aborder et la méthode TPOT est très bien expliqué avec des études dans le domaine du médical.

References

Git de l'outil TPOT - http://epistasislab.github.io/tpot/

Article de towards data science - https://towardsdatascience.com/tpot-automated-machine-learning-in-python-4c063b3e5de9

Article de towards data science - https://towardsdatascience.com/tpot-pipelines-optimization-with-genetic-algorithms-56ec44ef6ede
Vidéo Youtube expliquant comment utiliser TPOT - https://www.youtube.com/watch?v=YzRpjfUCFsQ

Apprendre à utiliser TPOT sur Python - https://www.datacamp.com/community/tutorials/tpot-machine-learning-python