Stage INFO2: Implémentation et Évaluation d'un algorithme d'apprentissage automatique pour la découverte automatique de connaissances biologiques

1er juin 2017

1 Contexte

Le processus de découverte scientifique en biologie fonctionne principalement de façon empirique et itérative. Á partir de l'observation de phénomènes au cours de multiples expérimentations, un biologiste a pour objectif de construire un modèle théorique cohérent avec les observations expérimentales, sous la forme de principes et de lois. Un modèle est ainsi valide tant qu'aucune observation expérimentale ne contredit ses lois, et une révision du modèle doit être envisagée à chaque fois que cela survient, de sorte que de plus en plus d'observations soient explicables par ce modèle.

En pratique, ce processus pose de nombreux problèmes. Sur le plan pratique tout d'abord, alors que le contrôle rigoureux des expériences et leur reproductibilité est fondamental pour s'assurer qu'une observation expérimentale est réelle et non issue d'une erreur de manipulation ou de protocole, la majorité des expériences biologiques réalisées aujourd'hui est difficile à reproduire. Sur un plan plus théorique ensuite, le choix d'une révision d'un modèle à partir d'observations inexpliquées est difficile, car le nombre de possibilités est très important et qu'il n'est pas toujours évident de faire un choix objectif entre elles. Enfin, le choix des expériences à réaliser pour améliorer au plus vite le modèle est une question importante pour bénéficier au plus vite de découvertes scientifiques majeures en réduisant les coûts nécessaires ; or, cette considération est encore très peu étudiée dans la communauté bioinformatique.

Un robot scientifique est une machine reproduisant de façon automatique le processus de découverte scientifique, capable de choisir, exécuter et analyser des expériences de façon itérative, grâce à des technologies de pointe en intelligence artificielle. Utiliser ce

type de robot présenterait idéalement de nombreux avantages tels que l'assurance d'expériences reproductibles, exécutées dans des environnements contrôlés, ainsi que l'utilisation de méthodes objectives à la fois pour la révision des modèles théoriques et le choix des expériences pour les itérations suivantes. Néanmoins, ces questions sont encore majoritairement des problèmes ouverts de recherche.

Le projet européen Adalab (Adaptive Automated Scientific Laboratory) [1] a pour but de repousser les limites liées à la découverte (semi)-automatisée de connaissances par des équipes constituées à la fois de scientifiques humains et de robots. Les enjeux majeurs sont une plus grande autonomie pour les robots scientifiques ainsi qu'une meilleure interface entre humains et machines, pour une meilleure répartition des tâches.

Ce projet touche de nombreux domaines en technologies de l'information dont l'apprentissage automatique (machine learning) et la représentation de connaissances. Le cadre applicatif de ce projet porte sur des problématiques très actuelles en biologie cellulaire ayant un impact fort sur la compréhension du cancer et du vieillissement.

L'équipe A^3 du Laboratoire d'Informatique de Paris Nord (LIPN) est un acteur de ce projet, notamment sur les problématiques d'apprentissage à partir de données, de révision de modèle et de raisonnement automatique.

2 Sujet de stage

Actuellement, l'équipe A^3 a conçu un algorithme d'apprentissage automatique de modèles de régulation biologiques à partir de données temporelles, basée sur l'agrégation de composants très simples (des arborescences) de façon à obtenir un modèle consensuel, robuste à des variations mineures dans les données utilisées. Cet algorithme présente actuellement de bonnes performances sur des données biologiques "benchmark" dans le domaine, mais peut être sensiblement amélioré.

Le but de ce stage sera de comprendre, d'implémenter, d'intégrer à l'existant et d'évaluer une méthode de l'état de l'art [2] en apprentissage automatique, mêlant méthodes statistiques et théorie des graphes, basée sur des structures graphiques plus complexes que les arborescences de l'algorithme A^3 : les modèles à largeur arborescente bornée. Ces modèles devraient permettre d'étendre de façon naturelle les contraintes de l'algorithme actuellement en place et d'améliorer ses performances sur les données considérées.

La différence de performances entre les versions de l'algorithme devra être évaluée sur les données d'un challenge récent en bioinformatique proposant des données temporelles pour l'apprentissage de réseaux de régulation : HPN DREAM 8 : [2]. L'objectif

1. http://www.adalab.mib.manchester.ac.uk/
2. https://www.synapse.org/#!Synapse:syn1720047/wiki/55342

est d'atteindre les premières places de ce classement et ainsi améliorer sensiblement le classement actuellement atteint.

3 Compétences

Ce stage nécessite des connaissances en programmation, en théorie des graphes et en probabilités. Une connaissance de R est un atout supplémentaire pour le choix du candidat. Pour la partie analyse des résultats, une connaissance d'outils de reporting, tels que knitr ou Shiny en R, ou HTML/Javascript, est aussi un plus.

4 Contact

— Anthony Coutant : anthony.coutant@lipn.univ-paris13.fr

Références

[1] Nie, S., Maua, D. D., de Campos, C. P. and Ji., Q. *Advances in learning Bayesian networks of bounded treewidth.* Advances in Neural Information Processing Systems, 2285–2293, 2014.

[2] Nie, S., de Campos, C. P., Ji, Q. *Learning Bayesian networks with bounded tree-width via guided search.* Thirtieth AAAI Conference on Artificial Intelligence, March 2016.