Méthodes de prédiction du risque d'exacerbation dans des maladies inflammatoires chroniques à partir de données hétérogènes

Directeur : Macha Nikolski
Co-directeur : Thierry Schaeverbeke
mail : macha.nikolski@labri.fr
Équipe : Modèles et Algorithmes pour la Bioinformatique et Visualisation du LaBRI

Mots-clé : machine learning, intégration de données, méthodes prédictives, maladies inflammatoires

Résumé

Dans cette thèse on s'intéressera au développement de méthodes de machine learning afin de déterminer les combinaisons optimales les prédicteurs du risque d'exacerbation de plusieurs maladies inflammatoires (asthme, polyarthrite rhumatoïde, ...), intégrant un ensemble de données hétérogènes de grand volume provenant de l'environnement immédiat, du microenvironnement pulmonaire et des symptômes perçus par le patient.

Contexte

Les maladies inflammatoires chroniques représentent la 3e cause de mortalité dans les pays développés, après le cancer et les maladies cardiovasculaires ; elles sont responsables d'une morbidité importante et ont un coût sociétal considérable. Leur prévalence est en constante augmentation dans les pays développés.

L'originalité de ce projet est de développer une stratégie d'analyse de type « big data » multifactorielle pour évaluer les risques d'exacerbations hétérogènes dans des maladies complexes où des facteurs d'environnement influent sur le tractus respiratoire.

Ce projet fait partie intégrante du FHU ACRONIM, porté par le CHU et l'Université de Bordeaux et de l'ATT G2P porté par le Département Sciences de la Vie. L'accès privilégié aux données collectées dans le cadre du FHU ACRONIM permettra un développement méthodologique majeur allant des développements algorithmiques jusqu'à leur passage à l'échelle et déploiement grandeur nature. Une telle validation de la méthodologie servira de point de référence dans le domaine.

Objectifs de la thèse

Développement de méthodes de recherche de combinaisons optimales de marqueurs prédicteurs du risque d'exacerbation de plusieurs maladies inflammatoires (asthme, polyarthrite rhumatoïde, ...), intégrant un ensemble de données hétérogènes de grand volume provenant de l'environnement immédiat, du microenvironnement pulmonaire et des symptômes perçus par le patient.
Le succès de ce projet sera conditionné par (i) l'établissement d'une représentation normalisée et unifiée de donnés faisant appel à des référentiels de termes internationaux, (ii) le développement de programmes et d'algorithmes de collecte de données

longitudinales issues de bases de données publiques (facteurs environnementaux), (iii) la mise en place de méthodes d'apprentissage supervisées capable de traiter de grands volumes de données longitudinales.

Dans un premier temps, le candidat devra mettre en place des analyses primaires de données (normalisation, assignations taxonomiques etc). Pour cela il pourra s'appuyer sur les équipes cliniques et biologiques afin de garantir que toutes les données qui seront utilisées dans le projet soient d'une part cohérentes et d'autre part représentent de façon fidèle les phénomènes biologiques et cliniques d'intérêt.

Dans un deuxième temps, le candidat étudiera l'application des algorithmes classiques du type apprentissage supervisé de machine learning (SVM, Random Forest etc) aux données du projet. L'apprentissage supervisé est un domaine de l'informatique et des statistiques où l'on s'intéresse à la construction automatique d'un modèle capable de faire des prédictions. En général, cela implique un algorithme de machine learning qui déduit certaines propriétés d'un ensemble de données d'apprentissage. Cet ensemble est composé d'instances (ou items), c'est-à-dire d'objets associés à un ou plusieurs attributs (correspondant à la valeur associée à une caractéristique). Dans ce projet, on s'intéresse aux problèmes de la classification : (1) feature selection qui s'intéresse à la recherche des paramètres discriminants et (2) la construction d'un classifieur capable de prédire les attributs d'intérêt (dans notre cas le risque d'exacerbation) à partir de ces descripteurs. Le candidat devra pour cela proposer de nouvelles méthodes d'analyse permettant de combiner les données qualitatives (remontées par le patients), quantitatives (capteurs environnementaux), cliniques et de séquences métagénomiques. La fusion de ces données hétérogènes nécessitera le développement de méthodes spécifiques en sémantique / ontologies et de fusion d'information. Pour ce faire, le candidat pourra profiter de l'expertise du LaBRI dans le domaine ainsi que celle de nos collaborateurs de Capgemini.

Enfin, suite à l'évaluation de la précision des prédictions obtenues dans la première étape, nous nous intéresserons à la mise en place d'un classifieur « presque optimal » basé sur le deep learning. En effet, le deep learning tente de modéliser les données en utilisant de modèles composés de multiples transformations non-linéaires. Une des promesses de cette méthodologie est de remplacer des descripteurs construits sur mesure par des descripteurs extraits automatiquement des données et pouvant donc évoluer au fur et à mesure de l'acquisition de celles-ci.

La répartition de l'encadrement :

- *Macha Nikolski* : analyse bioinformatique et biostatistique des données, élaboration des algorithmes de prédiction
- *Thierry Schaeverbeke* : mise en place des cohortes, des protocoles de suivi des patients, questionnaires d'évaluation cliniques, analyse des facteurs d'environnement