Analyse statistique de données issues du traitement de la parole pour aider au diagnostic de la maladie de Parkinson

Explorer de nouvelles voies pour diagnostiquer la maladie de Parkinson est la motivation centrale du projet pluridisciplinaire dans lequel ce stage s'insère. Plus précisément, l'analyse de la parole est susceptible de permettre de détecter la présence de la maladie de Parkinson mais également faire la différence entre Parkinson (PD) et l'atrophie multi systématisée (MSA). Ces deux maladies dans leur stage le plus précoce présentent des symptômes identiques. Or la prise médicamenteuse destinée à Parkinson utilisée sur MSA dégrade très rapidement l'espérance de vie du patient. Il est important de s'assurer du bon diagnostic. L'analyse de la parole permettrait d'apporter des informations pertinentes à cet effet. Ce stage se situe dans le cadre du projet Voice4PD-MSA financé par l'ANR et permettant une collaboration entre l'Université Paul Sabatier (laboratoires IMT et IRIT), les CHU de Bordeaux et Toulouse et les équipes de recherche de l'INRIA Bordeaux. Afin de pouvoir commencer les analyses de la parole, nous souhaitons débuter une étude sur la détection de la maladie de Parkinson et nous allons collaborer avec le CHU d'Aix-en-Provence et du laboratoire LPL et travailler sur un corpus existant de 273 patients.

Les données à analyser sont issues d'une cohorte de plusieurs centaines de patients répartis en 1/ patients pour lesquels un diagnostic de maladie de Parkinson a été posé et 2/ d'autres patients-témoins non atteints par la maladie. Les informations dont on dispose pour ces patients sont d'une part, d'ordre clinique (âge, antécédent) et d'autre part de caractéristiques vocales mesurées suite à un enregistrement de la voyelle /a/ tenue. De nombreux paramètres peuvent être extraits de la parole dérivés des paramètres spectraux, de la fréquence fondamentale et des formants.

Les analyses statistiques à mener sur ces données aborderont d'abord une démarche non supervisée (c'est à dire sans tenir compte de l'état du patient malade / témoin). Cette démarche, basée sur des analyses exploratoires multidimensionnelles (Analyse en Composantes Principales, Analyse des Correspondances, classification hiérarchique) visera à comprendre la structure des données et à envisager une stratégie de sélection de variables (les variables attendues suite au traitement de la voie pourront être très nombreuses – plusieurs milliers – selon les traitements effectués). Dans un second temps, des analyses supervisées, intégrant ou pas des stratégies de sélection de variables, viseront à prédire au mieux le statut des patients. Dans cet objectif, les méthodes discriminantes linéaires (Analyse Factorielle Discriminante, PLS-DA) ainsi que des techniques d'apprentissages (Random forest, Support Vector Machine...) seront mises en œuvre. Ces différentes approches s'intégreront dans une démarche globale de validation croisée afin de quantifier au mieux les résultats des méthodes quant à la prédiction du statut des patients.

Durée : 4 à 6 mois selon disponibilités.

Formation : informatique, ingénierie mathématique, statistique.

Niveau : M1-M2.

Pré-requis : connaissance des langages R ou Python.

Contact : Sébastien Déjean (sebastien.dejean@math.univ-toulouse.fr), Laurent Risser (IMT), Jérôme Farinas (jerome.farinas@irit.fr), Julie Mauclair, Oriol Pont, Etienne Sicard (IRIT)