Proposition de recherche doctorale

Machine learning from multimodal genetic and neuroimaging data for personalized medicine

Mots clés :
- Directeur de thèse : Olivier COLLIOT
- Co-encadrant(s) :
- Unité de recherche : Centre de recherche de l'Institut du cerveau et de la moelle épinière
- Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)
Personalized medicine aims at tailoring medical decisions, prevention and therapies to individual patients, based on their predicted risk of disease, evolution and response. In this approach, patients are characterized using rich multimodal measurements (genomics, medical imaging, biomarkers...). A central challenge is then to develop predictive models from these measurements. To that end, it is necessary to design new machine learning approaches that can fully exploit the different types of data. Neurodegenerative diseases, such as Alzheimer’s disease and Parkinson’s disease, are complex multifactorial diseases that represent major public health issues. In the context of these brain disorders, two types of data play a major role: genetics and neuroimaging. Genetics allow identifying factors that modulate the risk of a given disease, its evolution and response to treatment. It involves measurement of increasing complexity, from series of Single Nucleotide Polymorphisms (SNPs) provided by microarrays to high-throughput sequencing approaches such as whole-exome or even whole-genome sequencing. Neuroimaging allows measuring, in the living patient, different types of anatomical and functional alterations, using a variety of imaging modalities: anatomical, functional and diffusion magnetic resonance imaging (MRI) and positron emission tomography (PET). These two technologies have witnessed considerable development during the past 15 years. In the meantime, important advances have been made for processing and statistical analysis of these complex data. In particular, our laboratory has developed advanced machine learning approaches for disease prediction from neuroimaging data (Cuñénet et al., 2013; Gerardin et al., 2009). However, machine learning approaches that can adequately integrate neuroimaging and genetic data are currently lacking. The development of such approaches is particularly timely because massive datasets of patients with both imaging and genetic data are now available. One can cite for instance the Alzheimer’s Disease Neuroimaging Initiative (ADNI, http://www.adni-info.org/), the UK Biobank (http://www.ukbiobank.ac.uk/), the MEMENTO national cohort (http://www.ukbiobank.ac.uk/) or the Parkinson’s disease Progression Markers Initiative (PPMI, http://www.ppmi-info.org/). Methodological developments are challenging because of: i) the high dimensionality of both types of data (around $10^5$-$10^6$); ii) the complex multivariate interactions between variables, i.e. variables usually have only a mild effect when considered in isolation and only their combination can result in higher predictive power; iii) the structure of these data (spatial and anatomical structure for brain images, genomic structure) that needs to be adequately modeled.

{{Research program}} The objective of this PhD thesis is to develop and validate new statistical learning approaches that can integrate genetic and neuroimaging data, in the context of personalized medicine for neurodegenerative diseases. The main strategy we propose to pursue is to adequately model the structure of both genetic and neuroimaging data. Such strategy aims at constraining learning procedures to better handle the high-dimensionality and at providing interpretable results. Neuroimaging data is structured by the geometry of anatomical structures, their relations and their connectivity. Genetic data is also highly structured: variants are grouped within genes, their dependency is structured by genomic architecture, and genes interact within pathways. Our team has recently proposed new approaches for integrating the structure of neuroimaging data into statistical learning approaches (Cuñénet et al., 2013). In the context of the present thesis, we will focus on genetic data and its integration with neuroimaging. First, we propose to use the grouping of variants into genes and that of genes into common pathways and to select only relevant groups of genes/pathways. For that, we propose to use combinations of l2 and l1 norms, in the spirit of the group lasso approach (Yuan and Lin, 2006). We then propose to take into account the interactions of genes within a given pathway. Such interactions can be modeled using a graph which defines new regularization operators that can be introduced within the learning process through the definition of a new kernel (Kondor and Lafferty, 2002). We then propose to integrate imaging and genetic data through the definition of new kernels. We will define new kernels and similarity measures for genetic data. We will combine them with kernels for imaging data, using for instance multiple kernel learning (Gönen and Alpayd?n, 2011). We propose to introduce specific constraints on the topographical expression of genes. Indeed, genes present with a differential expression in different brain regions. Such information appears particularly useful for integration with neuroimaging data. We will thus aim to integrate this information in the definition of new kernels. The methodological developments of this PhD thesis will be prototyped and validated using the ADNI (Alzheimer’s Disease Neuroimaging Initiative) database which is publicly available. Overall, it comprises over 300 patients with AD, over 850 patients with mild cognitive impairment (MCI) and over 350 control subjects. Neuroimaging data includes anatomical MRI, resting-state functional MRI, diffusion MRI, amyloid and fluorodeoxyglucose (FDG) PET. Genetic data consists in both micro-array data covering 620,901 SNPs and whole genome sequencing data. Long term follow-ups are available. Our team has extensive experience with preprocessing of this dataset. The PhD candidate will thus have access to all the necessary support for performing this experimental part. Furthermore, our team has close collaborations with the bioinformatics team of the ICM and with specialists of genetics of neurodegenerative disorders at ICM (A. Brice, I. Le Ber). The PhD student will benefit from these collaborations, in particular for processing of genetic data and the interpretation of results.


Résumé du projet de recherche (Langue 2)

This PhD thesis aims at developing new machine learning approaches that can integrate genetic and neuroimaging data. The main scientific challenge is of methodological nature and will be to integrate, in a coherent framework, heterogeneous and high dimensional data coming from different modalities. In particular, both genetic and neuroimaging data have specific structures that need to be adequately taken into account: spatial and anatomical structure, interaction of variants and genes within pathways. A major challenge will thus be to adequately integrate these different structures into a coherent statistical learning framework.

Informations complémentaires (Langue 1)

We currently have collaborations in the field of statistical learning for neurodegenerative disorders with the teams of Sebastien Ourselin and Daniel Alexander at University College London (UK). The work of the PhD candidate could benefit from these collaborations through participation to joint scientific meetings and possible visits to London.

Informations complémentaires (Langue 2)

Our work could also benefit from collaborations with the research team of Sebastien Ourselin and Daniel Alexander at University College London (UK). The work of the PhD candidate could benefit from these collaborations through participation to joint scientific meetings and possible visits to London.
The PhD thesis will be conducted within the ARAMIS team (www.aramislab.fr) at the Brain and Spine Institute. ARAMIS is a joint team between CNRS, INRIA, Inserm and University Pierre et Marie Curie. The candidate should have a strong background in machine learning, with a specific emphasis on statistical aspects. Previous experience with brain imaging or genetic data is not necessary.